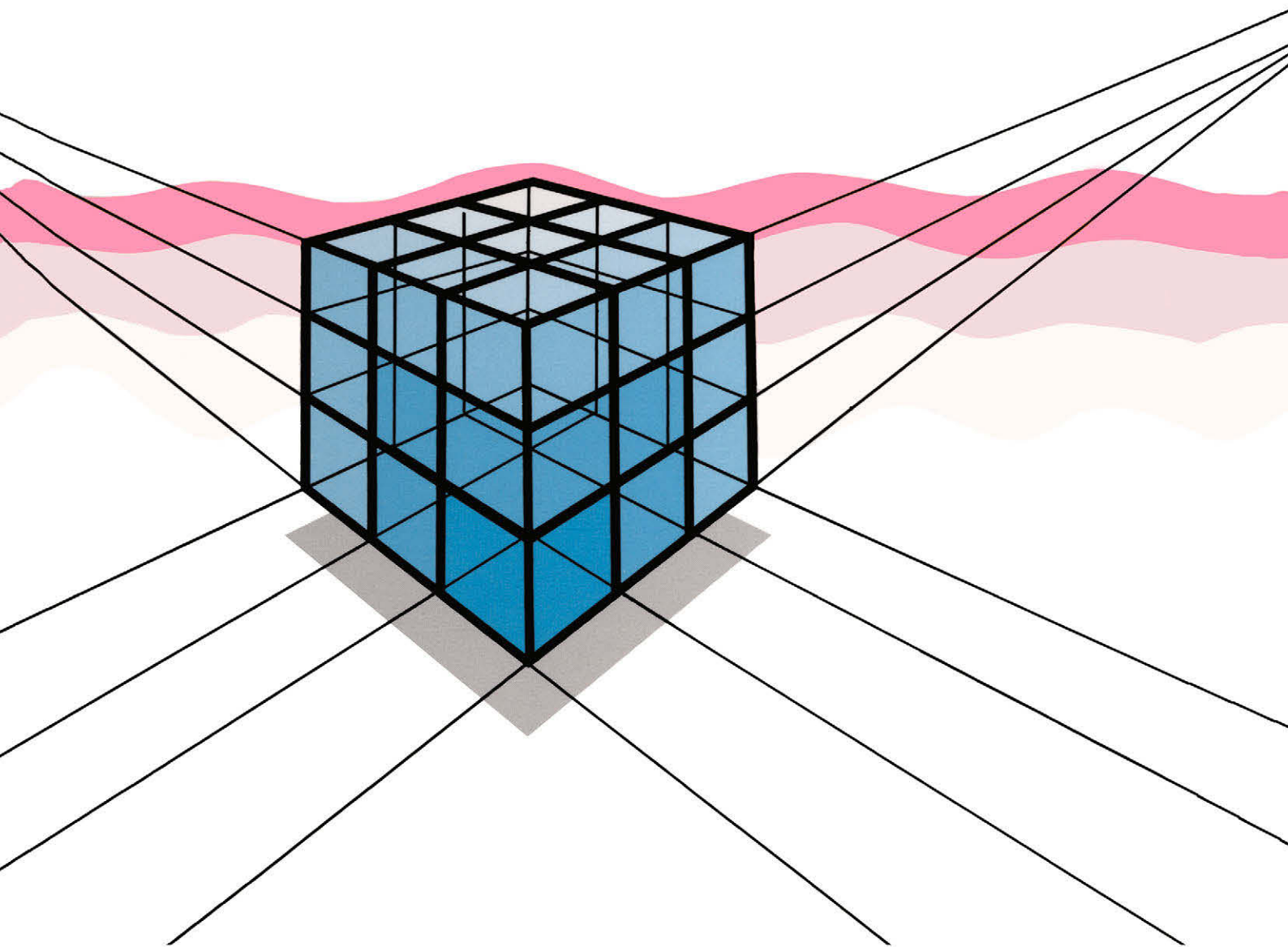
STUDIES IN COGNITIVE SYSTEMS

INTERNAL AFFAIRS

Making Room for Psychosemantic Internalism

by

KEITH L. BUTLER



INTERNAL AFFAIRS

STUDIES IN COGNITIVE SYSTEMS

VOLUME 21

EDITOR

James H. Fetzer, *University of Minnesota, Duluth*

ADVISORY EDITORIAL BOARD

Fred Dretske, *Stanford University*

Charles E.M. Dunlop, *University of Michigan, Flint*

Ellery Eells, *University of Wisconsin, Madison*

Alick Elithorn, *Royal Free Hospital, London*

Jerry Fodor, *Rutgers University*

Alvin Goldman, *University of Arizona*

Jaakko Hintikka, *Boston University*

Frank Keil, *Cornell University*

William Rapaport, *State University of New York at Buffalo*

Barry Richards, *Imperial College, London*

Stephen Stich, *Rutgers University*

Lucia Vaina, *Boston University*

Terry Winograd, *Stanford University*

INTERNAL AFFAIRS

Making Room for Psychosemantic Internalism

by

KEITH L. BUTLER

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-90-481-5104-2 ISBN 978-94-017-1921-6 (eBook)
DOI 10.1007/978-94-017-1921-6

Printed on acid-free paper

All Rights Reserved
© 1998 Springer Science+Business Media Dordrecht
Originally published by Kluwer Academic Publishers in 1998
No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the copyright owner.

Pour

Mireille

SERIES PREFACE

This series will include monographs and collections of studies devoted to the investigation and exploration of knowledge, information, and data-processing systems of all kinds, no matter whether human, (other) animal, or machine. Its scope is intended to span the full range of interests from classical problems in the philosophy of mind and philosophical psychology through issues in cognitive psychology and sociobiology (regarding the mental abilities of other species) to ideas related to artificial intelligence and computer science. While emphasis will be placed upon theoretical, conceptual, and epistemological aspects of these problems and domains, empirical, experimental, and methodological studies will also appear from time to time.

A fundamental but still unresolved issue in the philosophy of mind remains the internalism/externalism debate over the nature of the contents of mental states. In this volume, Keith Butler elaborates three important but not uncontroversial claims, specifically: that mental states must be individuated on the basis of their contents; that they typically refer to external states (and are therefore "broad" or "wide"); yet mental contents also supervene upon the internal states of individuals. He thereby avoids the common pitfalls of assuming that broad content presupposes or implies externalism and that internalism presupposes or implies mental states that make no reference to external states. The result is an illuminating study that not only clarifies but even promises to solve this difficult problem.

J. H. F.

CONTENTS

Preface xi

Introduction 1
1. Preliminary Remarks 1
2. The Argument 5

Part I: Externalism and the Propositional Attitudes

Chapter 1: Externalism and Twin Earth 12
1. Preliminary Remarks 12
2. Twins 12
3. A Fantasy 19
4. Concluding Remarks 24

Chapter 2: Self-Knowledge 25
1. Preliminary Remarks 25
2. Externalism and Knowledge of Content 28
3. Internalism and Knowledge of Content 57
4. Concluding Remarks 63

Chapter 3: Skepticism 64
1. Preliminary Remarks 64
2. The QB Strategy 67
3. Criticism of the QB Strategy 74
4. Beyond the QB Strategy 80
5. Concluding Remarks 82

Part II: Anti-Individualism and the Cognitive Sciences

Chapter 4: Computational Vision Theory 84
1. Preliminary Remarks 84
2. What Is Individualism? 86
3. Burge's Argument 92
4. Shapiro's Argument 103
5. Epistemic and Metaphysical Projects 106
6. Segal's Liberal Strategy Reconsidered 119
7. The Conservative Defense Strategy 124
8. Concluding Remarks 130

Chapter 5: Causal Powers 131
1. Preliminary Remarks 131
2. A Prima Facie Case Against Anti-Individualism 132

3. The Argument From Scientific Taxonomy 144
4. The Argument From Behavior 162
5. Concluding Remarks 174

Chapter 6: Cognitive Explanation 177
1. Preliminary Remarks 177
2. What Are Cognitive Systems? Some Preliminary Thoughts 178
3. Explanatory Virtues 182
4. Systemic Anti-individualism 199
5. Concluding Remarks 223

Notes 227

References 247

Index 257

PREFACE

This book is not written for posterity. It is meant to constitute part of my contribution to a continuing debate at the intersection of the philosophy of mind, the philosophy of language, epistemology, and the philosophy of cognitive science. The debate is over how mental states are individuated. Many philosophers, call them externalists (or anti-individualists), believe that the individuation of mental states requires appeal to an individual's social and/or physical environment. I am not among them; I am an internalist (or individualist). This book attempts to show that the leading proponents of externalism make a lot of mistakes in defending their view. This is either because the view is false, so any defense of it is bound to make some mistake or another, or it is because the people defending externalism have not been very careful, or both.

I have included some introductory material so that those not already familiar with the issues involved might nevertheless gain entry into the debate. The intended audience, however, consists in those professional philosophers and students of philosophy who are already familiar with and interested in the question of how mental states are individuated. I do not spend much time trying to motivate an interest in this issue. Like all philosophy, it is interesting and worthwhile if one is blessed (or cursed) with the appropriate kind of intellectual curiosity. Those who do not meet this requirement proceed at their own risk.

This book is a deliberate antidote to four horrible tendencies in recent philosophical writing. The first tendency is found in the work of those writers who attempt to imitate the style of philosophers like Fodor, Dennett, and Paul Churchland. Whereas Fodor, Dennett, and Churchland have the skill and intelligence to frame their views in clever and interesting language, those whose writing they have influenced lack either the writing skill of a Fodor or a Dennett or a Churchland, or the intelligence, or both. As a result, there is a large amount of philosophical writing that is cloying rather than clever. Moreover, in their effort to be clever, these writers get sloppy; they miss distinctions and obscure issues that are difficult enough already without their unhelpful contributions; see Andy Clark's recent book for an example of this tendency. I do not pretend to be clever, but I do claim to be clear; the writing in this book is as straightforward as it gets.

The second tendency I have in mind is found in the work of those philosophers who seem to want to be scientists (usually neuroscientists, or cognitive scientists -- sometimes even cell biologists or some such thing) rather than philosophers. These writers have produced volumes of science journalism, and very little of philosophical interest. Such

writers often claim that the study of mind should be an interdisciplinary affair, and that the distinction between philosophy and theoretical science is of no interest to them; they are interested in the mind (or mind/brain, as they sometimes call it), not turf wars between disciplines. This pedestrian pseudo-insight, however, makes it seem that such writers are at least producing something of scientific interest, even if it is of little or no philosophical interest. But this is simply false. The writers I have in mind have, at best, only cursory training in the relevant sciences, and certainly much less than the average graduate student of neurophysiology or computer science. They seem to want to participate in the respect that true scientists have earned without doing the hard work necessary to be trained properly in one of the sciences at issue. Their work is of little interest to scientists because it usually consists in very simplistic summaries of real scientific work; their work is of little interest to philosophers because it is devoid of philosophy. I know of one such writer, whose initials are Bill Bechtel, who has written introductions to some science or another for philosophers, and introductions to some area of philosophy or another for scientists; anything, it would seem, to avoid having to do any real science or real philosophy! When these journalists of science do try to draw philosophical lessons from their reports, they tend to reveal their philosophical ignorance. It is hard to know whether such philosophy is so naive because the person neglected her philosophical education in order to become a non-expert in the relevant sciences, or because the person is a poor philosopher trying to cover up her weaknesses with an impressive-sounding scientific vocabulary. We do know that such writers try to blind philosophers with science and scientists with philosophy; it is hard to know how else some of this so-called work managed to get through the reviewing process. The result, at any rate, is neither science nor philosophy; and it does a real disservice to those, such as Fodor, Churchland, Fetzer, and others, who do responsible interdisciplinary work. I expose some of this bad mix of science and philosophy (but certainly not all) in the latter half of this book (particularly in the last chapter). When I appeal to the sciences in this book, I do so to make a clear point of direct relevance to a philosophical issue, or to show that the appeals that others have made to the sciences are confused in one way or another.

The third tendency I have in mind is the tendency of philosophers to write books that are direct responses to the work of a limited number of other philosophers, and then try to pass the book off as a general reaction to broad currents of philosophical thought without identifying precisely whose work they are reacting to at any given moment. This procedure has resulted in books that are vague on important details, or overstated in their purported significance. The

cause of these defects, the failure to identify one's targets clearly, is little more than laziness. In this book, I have been careful to direct my remarks to specific works of specific philosophers whose work is placed in the context of clearly articulated currents of philosophical thought. The book that has resulted from this hard work is quite clear, though it cannot be read quickly or cursorily. It is, at times, when detailed responses are presented, difficult to read; but it is only as difficult as the details require. This is a small price to pay for intellectual honesty.

The fourth tendency I have in mind is not unrelated to the previous one; it is the tendency of some philosophers to ignore the work of all but a few leading philosophers. This is easy to do in the information age, where relevant work proliferates like a crafty virus. But no one ever said scholarship is easy. Though each person should be permitted to develop her views as she sees fit, the fact is that a lot of philosophers are, as it were, reinventing the wheel. No one can track down every relevant piece of work; this is to be expected when new work appears every day. But some effort should be made to read the leading journals to see what others are saying about the issues one is writing on. I have done as much of this as I was able to do in the time I had available to me in writing this book, and the result is a book that casts its net fairly wide. Though more relevant work has been published in the time between the completion of this book (in 1996) and its publication (in 1998), the book should provide fairly thorough coverage of the work available at a recent stage in the development of the debate between internalists and externalists.

There are several people I would like to thank for the help they have provided me in getting this book published. Above all, I would like to thank George Harris, Paul Davies, and Jim Fetzer. Each one provided me with tangible encouragement and assistance at various stages in the difficult process of finding a suitable outlet for this book; I will always be grateful for their confidence and friendship. I would also like to thank Norton Nelkin, whose influence on my thinking is apparent on every page. It was an honor to have known him.

I would also like to thank those who have read and commented on all or parts of this manuscript: Tomoji Shogenji and Eric Saidel provided me with valuable comments on the entire manuscript at one stage of its development. Dana Nelkin posed a number of challenging questions about my views that resulted in immediate alterations to the manuscript. An anonymous reviewer for Kluwer Academic Press provided me with a number of challenging comments that resulted in alterations to the manuscript. The manuscript was used as the basis for a graduate seminar I taught in the spring of 1997; I received a number of interesting comments and suggestions from the students of that seminar. George Harris, Paul Davies, Carolyn Morillo, Jim Stone and Judy Crane

provided valuable commentary on certain parts of the manuscript. Gabe Segal, Jerry Fodor, Rob Wilson, Larry Shapiro, and many anonymous referees for many journals all read and commented on papers that eventually formed the bases for various chapters in the book.

Chapter 2 is a much expanded version of a paper entitled "Externalism, Internalism, and Knowledge of Content", which appeared in *Philosophy and Phenomenological Research.* Chapter 3 is a much expanded version of a paper entitled "Externalism and Skepticism", which appeared in *Dialogue: The Canadian Philosophical Review.* Chapter 4 is an expanded version of a paper entitled "Individualism and Marr's Computational Theory of Vision", which appeared in *Mind and Language.* Section 3.3 of chapter 4 is an expanded version of a paper entitled "Content, Computation, and Individuation in Vision Theory", which appeared in *Analysis.* Section 3.3 of chapter 5 is an expanded version of a paper entitled "Content, Causal Powers, and Contexts", which appeared in *Philosophy of Science.* Section 4 of chapter 6 is an expanded version of a paper entitled "The Scope of Psychology", which appeared in *PSA* 1994. I thank these journals for their permission to use this material.

I would also like to acknowledge my decidedly non-philosophical former colleagues in St. Louis. The many dim bulbs in the dark halls of that department concealed a woeful collection of wolves in cheap clothing. Were it not for their cruel incompetence and inexplicable stupidity I might never have had the good fortune of joining the bright and energetic Department of Philosophy at Western Washington University. For that I am grateful.

On a more personal note, I would like to thank Tempete and Otis, without whose generous attention this book would have been done several months earlier. And, finally, I would like to express a special debt of gratitude to my wife, Mireille, to whom this book is dedicated. To her I say: *Pour que tout le monde le sache, je voudrais dire que je t'aime. La meilleure affaire est l'affaire interne entre toi et moi.*

KLB
Western Washington University

INTRODUCTION

1. *Preliminary Remarks*

For all I know I might be dreaming right now. Or perhaps there is an evil Demon deploying all his powers to deceive me. Maybe, even, I am a brain in a vat. On the other hand, maybe things are just the way they seem to me. Some philosophers are inclined to say that these considerations put me in an epistemological quandary; I just do not know whether any of my beliefs about the external world are true, indeed, whether there is an external world at all. Through these sorts of arguments, Descartes is reputed to have left to the modern student of philosophy a legacy. It is not, however, (just) the specter of skepticism that we have inherited (nor, of course, did we inherit Descartes' substance dualism). Rather, it is the idea that, in the various scenarios sketched above, I might be having all the same thoughts, regardless of whether they are true or radically mistaken. This has been taken by some to suggest that my cognitive life is distinctively private and insulated from the rest of the world; the suggestion is that our cogitations are *internal affairs*, to be individuated independently of our environments.[1] We may call this Cartesian legacy *internalism* (or, for some purposes and in some contexts, *individualism*).

Internalism can be thought of as a thesis about the fundamental presuppositions of psychology. Whether one is a practicing psychologist, or a person on the street, one presupposes an inventory of thoughts to which appeal is made in the attempt to understand others' thinking and behavior (just as, in economics, shop owners and economists alike presuppose an inventory of currencies and denominations in trying to reach an understanding of economic trends). It is assumed almost universally that thoughts are categorized according to their content (i.e., what they are about), and the attitude one takes to the content (e.g., belief, desire, etc.). For example, my belief that water is wet is the thought that it is in virtue of its content (water is wet) and the attitude I take toward that content (belief). Had I instead taken an attitude of hoping toward that content, or had I instead taken an attitude of believing toward a different content, I would have been thinking a thought of a different kind.

Given this way of categorizing thoughts, a further question arises: Suppose that I am thinking the thought that water is wet. What is it in virtue of which I am, at this moment, thinking a thought with this particular content? This is not the question of what has *caused* me to think a thought with this content (perhaps my seeing a swimming pool caused me to have that thought). Rather, it is the question, *What facts must be true of me if I am to have a thought with that content* (regardless of how I might have been caused to have this

thought)? This question, in the jargon of the discipline, is a question of *individuation*. How are mental states individuated?

Since it has been clear for some time that our brains are intimately involved in our ability to think, a natural answer to this question, which I favor, is that I am thinking this thought due to the fact that my brain is in some particular state or configuration. We may not know *what* are the particular facts about my brain in virtue of which I have this thought, but, according to this view, it is nevertheless true *that* there are particular facts about my brain in virtue of which I have this thought. This is internalism; the facts in virtue of which I am having this thought are facts about what is occurring inside me.

Over the last 15 years or so, however, most philosophers have abandoned internalism in favor of *externalism* (sometimes also called *anti-individualism*).[2] Externalism holds that the state of my brain, while it might be an essential ingredient in determining that I am thinking at all, is insufficient to determine the thought that I am thinking as the thought that water is wet; I must, in addition, reside in an environment that contains water and/or individuals who use the word "water" to mean water. A little science fiction will help illustrate the externalists' claim: According to externalism, were I instead an inhabitant of an otherwise identical environment that had something in the lakes and streams that looks like water, but is in fact made of something other than H_2O, I would be thinking a thought with a different content, even if I would not be able to tell the difference between water and this other substance that only looks like water. Thus, externalism maintains, it is not a state of my brain in virtue of which I think that water is wet; rather, it is a fact about my environment (that it contains H_2O rather than that other stuff) in virtue of which I think that water is wet, even if that fact is entirely unknown to me.

The externalist claim, then, is that psychological explanations presuppose an inventory of thoughts that are categorized not by what is going on inside the heads of individuals, but by the often-hidden nature of the environment in which the individuals reside. Studying psychology, then, would not be a matter of looking at how thoughts arise from brain states (as much neuroscience and neuropsychology presupposes), or how (and why) brain activity causes us to behave; rather, it would be a matter of understanding the environments in which individuals are embedded, and their relationship to it.

It is my claim that this analysis of the metaphysical presuppositions of psychology, this analysis of its philosophical foundations, is deeply misguided. Sociology and ecology might harbor metaphysical commitments that require attention to environmental factors and relations, but psychology, I will suggest, does not. I will

argue that there are no good reasons to suppose that psychological explanations presuppose externalism, and that the suggestion that psychology does have these presuppositions is in conflict with other presuppositions of psychology (and other more basic sciences) that are very difficult to deny (even externalists would not deny them). Therefore, externalism, as a theory of the presuppositions of psychology, is quite implausible.

Nelkin (1997) has argued that the debate between internalists and externalists is an extension of the 17th and 18th century debate between Cartesian rationalism and empiricism. Externalists are empiricists insofar as they take certain concepts (even if not all concepts) to arise from experience; in particular, the externalist maintains that the possession of concepts with empirical application is possible only when individuals interact appropriately with their environments. Internalist are rationalists insofar as they embrace the idea that the possession of all concepts, even those with empirical application, requires no environmental interaction; concepts might be caused to be tokened, or triggered, by the environment, but the internalist/rationalist takes the contents of these concepts to be determined by aspects of an individual, conceived independently of her environment. Ideas or concepts may be innate, or developed through a process of maturation, or both, but in any case are not individuatively linked to the environments in which their hosts reside; we could, in principle (though perhaps not in actual fact), be thinking the same thoughts no matter what the world happens to be like.

Though internalism in the philosophy of mind partakes in Descartes' rationalism, and is in any case descended from his epistemological thought experiments, we did not, in my view, need Descartes' authority to arrive at our internalistic proclivities. Most non-philosophers, and most philosophers' pre-philosophical intuitions, favor internalism. I am *in here*, while the rest of the world is *out there*; that is why I have a particular perspective on the world, one that is not shared by anyone else. When it comes to categorizing my thoughts, then, the natural inclination is to assume that only features *in here* are relevant. In more technical terminology, information transmitted publicly must first be transduced before it can be processed by the various perceptual and cognitive systems that characterize my brain; and what I make of that information, how I process it, would seem to be determined by the nature of my brain. There are many varieties of perceptual systems and physical boundries that insulate my cognitive mechanisms from the surrounding environment. So, again, when considering how my states of mind might be individuated, it is natural to think that attention should be restricted to what lies inside my head; that is where all the psychological action is, so that is what is relevant

to the classification of the psychological action. My relations to the external world may affect the nature of my cognition *causally*, but those relations do not seem to affect how my mental states divide in types or kinds; they do not bear on the nature of my cognition *individuatively*.

In philosophy, the rejection of Cartesian substance dualism has, if anything, increased the *prima facie* plausibility of the internalist agenda (*pace* Wilson, in press); I behave the way I do because of intrinsic properties of my nervous system, not on account of the relations my nervous system bears to the external world. There are non-arbitrary physical boundries insulating the seat of my thoughts from the swirling and ever-changing environments that they are about. The categorization of thoughts, it is natural to suppose, should concern what is insulated from the environment.

This is not necessarily a reductivist view of psychology (cf. Burge, 1986c, p. 719). The internalist need not, and probably should not, maintain that the study of psychology should become the study of nervous systems. As Fodor has often pointed out, psychologists cannot benefit from examining the brain if they would not be in a position to recognize psychological phenomena as they come across them; moreover, the internal properties in virtue of which one exhibits psychological phenomena might be so wonderfully complex as to be unrecognizable at the level of brain wiring and chemistry. For reasons such as these, an internalist would be better advised to suppose that mental phenomena supervene, in some sense, on the neural.[3] But internalism does, in my view, spring from mechanistic metaphysical intuitions much like those that inspired Descartes. These intuitions are, I believe, quite common, perhaps common enough to deserve the honorific 'commonsense'. I do not know if those metaphysical prejudices can be sustained, but nor am I obliged to give them up until someone can demonstrate that they are inadequate to the facts of psychology. Philosophical intuitions, like barnacles clinging to a wharf, are entitled to stand their ground until the tide of argument wipes them away.[4] Externalists must therefore provide us with a view that commands our assent. It will be my business in this book to show that their position is not ultimately tenable.

This topic is among the most widely discussed and important topics in all of philosophy. Its importance stems from two sources: First, there is the intrinsic interest of the subject. It is often maintained that the central questions of philosophy focus on human beings and their place in nature. Since externalists commend a radical reconception of ourselves as essentially embodied and embedded organisms, and hence a rejection of the most revered tradition in philosophy (i.e., the internalist tradition that was the signature feature of Descartes' philosophy); philosophical interest in the topic could not be greater. Moreover, as I have suggested, the externalist position is out of

sympathy with our pre-philosophical intuitions about the mind; my thoughts are the kinds of thoughts they are because of what is going in inside my head. Nothing will generate interest as much as controversy, and this issue gives rise to plenty of controversy. The second source of interest in the issue is its relevance to direction of research in the cognitive sciences. It has been maintained (by e.g., Clark, 1996) that externalism lies at the heart of a revolution in the cognitive sciences. According to this movement, individuals must be understood in their social and environmental contexts if they are to be understood at all. I think this is all much ado about not very much. Once the issue of psychological state individuation is understood properly, we will see clearly that the direction of research in the cognitive sciences should remain unaffected by the resolution of the debate between internalism and externalism.

2. *The Argument*

Before turning to what this book will provide, I want to be perfectly explicit about what it will *not* provide: I will not here produce a positive internalistic account of mental content and cognition (though I do indulge in a little internalist 'fantasy' in chapter 1). Such an account would, of course, be of paramount interest. Some would even say that it would be more interesting than the project I will undertake herein. But a well-developed positive account will also be difficult to come by, and must address an enormous extant literature (e.g., the vast and growing literature on the semantics of belief ascription). For this reason it must be left for another occasion. For there is plenty to address on the current occasion.

My aim in this book is to take a first step toward breaking the grip that externalism holds on contemporary philosophy of mind. I aim to provide both motivation and reason to reject externalism; I intend for these motivations and reasons to be principled and defensible in terms of intuitions and fundamental views that even externalists share (or *should* share). I aim to show that, whatever the virtues externalist arguments appear to have, the conclusion of those arguments, the externalist thesis itself, cannot be true (or, at least, is very unlikely to be true). Does this show that internalism *is* true? No. It could turn out that *neither* view is tenable. It might be, in other words, that the question of individuation leads to a kind of *paradox*, where two mutually exclusive and exhaustive options appear to face insurmountable difficulties. So, while I might not be able to snatch internalist victory from the jaws of externalist defeat, I do intend to make sure that externalists cannot themselves emerge victorious. What I hope to show in this book, then, is this: There are problems with externalism and certain arguments for it that are sufficiently significant

so as to warrant and motivate the search for and development of an internalistic alternative to externalism.

In this section, I will present the form my argument will take. In subsequent chapters, I will add flesh to these bones, and defend the resulting argument against externalist criticisms.

2.1 Part I: Externalism and the Propositional Attitudes

My main argument against externalism about propositional attitudes, to which Part I of this book is devoted, is relatively simple and straightforward. After laying out the sort of externalist argument with which I will be concerned, I will argue that externalism has implausible epistemological implications that threaten its coherence and plausibility.

The first chapter surveys Twin-Earth-style arguments for externalism, those of Putnam and Burge, about ordinary propositional attitudes (e.g., beliefs, desires). I then present a small fantasy; a brief glimpse at what a positive internalistic proposal would look like. This fantasy is necessarily vague and short on details (as fantasies tend to be), but it should give one an idea how one could coherently be an internalist. Once these arguments and a brief response are on the table, I turn to an examination of the epistemological implications of their externalist conclusion.

The first implausible implication of externalism, discussed in Chapter 2, is that externalism cannot account plausibly for the privileged nature of self-knowledge. Burge insists that thoughts are individuated in part by their contents; that is, contents are essential features of thoughts, and play a part in determining that a token thought is the type of thought that it is. According to externalism, certain of an individual's thoughts are such that they have the content they do in part because of the way that individual is related to her physical and/or social environment; that is, being related to the environment in a particular way is essential for certain thoughts to have the contents that they do. But, for such a content, if the environment's being some specific way is essential to one's having a thought with that content, then one cannot know what one is thinking (i.e., the content of one's thought) without knowing first whether the environment is that way. If S is to know that she is thinking that water is wet, for example, she must know that there is water in her environment rather than, say, twater. If she does not know that her environment contains water, then she cannot be sure that her thought content involves the concept of water rather than a concept of twater. Thus, if externalism is true, knowledge of the content of one's thoughts would depend on first having knowledge of the environment. It is a widely held epistemological view, however, one that I and most externalists endorse, that knowledge of the contents

of one's thoughts is privileged. That is, it is direct (it does not involve inference), and it is *a priori* (it does not depend on first knowing empirical facts about the external world that could only be known *a posteriori*). The truth of externalism is incompatible with the direct and *a priori* nature of self-knowledge.

The second implausible implication, discussed in Chapter 3, is related to the first. One response to the self-knowledge problem is to maintain that self-knowledge is *a priori* after all, and that this is not inconsistent with externalism in the least. S knows in a direct and *a priori* way that she has a thought involving the concept of water. If knowing this entails that she knows that there is water in her environment, this is not because she must *first* know that there is water in her environment; rather, it is because she can infer that there is water in her environment on the basis of her knowledge that she is thinking that water is wet. Putnam (1981, ch. 1) parlays this sort of reasoning into a purportedly *a priori* refutation of skepticism. There has been much discussion of this argument; ultimately, I think it fails to show that externalism forms the basis of an *a priori* refutation of skepticism. Still, I maintain, externalism is inconsistent with skepticism; if externalism is right, then self-knowledge would be sufficient for knowledge of the world. Many externalists, however, are uncomfortable with the idea that knowledge of the contents of one's thoughts is sufficient for knowledge of the environment; I think their discomfort is well-placed. Just as externalism should not make self-knowledge *a posteriori*, it should not make knowledge of the environment *a priori*. For it would follow from this that skepticism is, and has been all along, not just false but incoherent. Whatever one thinks about the truth of skepticism, however, it strains philosophical credulity to suppose that Descartes' worries issued from confusion.

On the strength of these implausibilities, I conclude that externalism should be regarded with suspicion. As I indicated, the main argument is simple. The real work to be done comes in defending the claims that constitute its premises. Many philosophers have undertaken to show, in many clever and subtle ways, that externalism does not face the unwanted epistemological implications, and is therefore not threatened as my argument suggests. So, if I am to sustain my argument, I have no choice but to show that the myriad defenses of externalism along these lines are mistaken. The effort to expose these mistakes can sometimes be painstaking and tedious. But it is necessary. I believe strongly that it is important to look at what exactly the various defenders of externalism have actually said. For the problems with their arguments are often subtle and small. The larger and underlying issues, I believe, are relatively simple and straightforward. But they are obscured by the shrewdness of externalist apologies. To

make the case that externalism is in trouble, then, we must move from the large to the small, from the manifest epistemological implausibilities of externalism to the devious ways in which they are covered up and distorted. It is only at this level of detail that the real troubles of externalism about propositional attitudes can be exposed and appreciated. Some may not want to work through the issues at this level of depth. There is, however, no other way to locate the troubles with the externalist apologies. Exposing those troubles for appreciation is the task of Part I.

2.2 Part II: Externalism and the Cognitive Sciences

If I am right that externalism about propositional attitudes is threatened by its robust epistemological implications, then there will be independent reason to look favorably upon internalist diagnoses of externalist arguments, as well as internalist accounts of the mental. But there remain battlefronts, involving issues in the cognitive sciences, on which externalism would not be threatened. In Part II, I consolidate the hard-fought gains of Part I by blocking several attempts to establish the scientific integrity of externalism.

The first issue, discussed in Chapter 4, concerns Burge's (1986a) argument that computational psychology, particularly as it is manifested in the theory of vision, is externalist. A careful examination of an extant computational theory in psychology, Marr's theory of vision, reveals that it is committed to an externalist individuation of visual states. Given the prominence of computational psychology, and the pivotal role that vision plays in psychology, a satisfying internalism cannot cede this territory to the externalist. In Chapter 4, I argue that the balance of reason and evidence does not favor externalism. Nothing in Marr's discussion of the content of visual states supports an externalist interpretation of the theory, and Twin-Earth-style thought experiments involving the application of Marr's theory suggest nothing about the individuation of visual states. To the extent that Burge is right that the results of an investigation of the individuation of visual states is suggestive of the individuation of psychological states generally, there is no reason to think that psychological states generally are individuated externally. This discussion will require careful consideration of the nature of internalism, as well as the arguments Burge deploys against it.

The second issue, discussed in Chapter 5, concerns the extent to which externalism is committed to an implausible view of mental causality. It is widely supposed that mental states are causally implicated in explanations of intentional behavior. One way to understand this is to suppose that the intentional properties in virtue of which mental states are mental (i.e., contents) supervene locally on the

properties in virtue of which mental states causally interact with each other and with perceptual (and proprioceptive, etc.) inputs and mechanisms governing behavioral outputs. This understanding of mental causality, which is heavily influenced by the computational paradigm in cognitive psychology, explains how mental states can interact causally in ways that are also in conformity with rationality and other forms of semantic coherence. I will consider certain recent refinements of this notion of supervenient causality that serve to minimize the extent to which mental causality is somehow suspicious or mysterious. I do not believe that this is the only way to make sense of the mutual satisfaction of causal and rational/semantic constraints, but it is the only one that has any precedent in science. No sciences count higher level properties as causal powers unless they bear certain well-specified relations to local properties of the causing system. If externalism is true, then either intentional properties do not bear such relations to the causally efficacious properties, or the causally efficacious properties are not internal properties of the behaving system. Either result would threaten the causal status of mental states and the extent to which psychological theories fall within the purview of science.

The third issue, discussed in Chapter 6, is related to the second. Some externalists have argued that behavioral explanations that appeal to properties that do not supervene locally on internal states of individuals are rich in explanatory power. Indeed, it is argued, the relational properties appealed to in these explanations are crucial to their success. Insofar as these explanations are successful, there is no reason why an externalist psychology should fall into scientific ill-repute. One variety of argument along these lines appeals to the fact that intentional behavior, as it is normally described, cannot be explained by internal factors alone; ordinary behavioral descriptions, as Wilson (1994) contends, are theoretically inappropriate to a purely internalistic psychology. Another variety of argument along these lines (e.g., Haugeland, 1994) questions whether a subject's physical boundary constitutes an explanatorily relevant boundary. I contend that these arguments misconstrue, in various ways, what a psychological explanation should be expected to explain. As a result, all sorts of non-psychological factors are wrongly taken to be psychological.

The substance of this book consists in various attacks on a wide array of externalist arguments. In spite of this negative drift, my hope is that these offensives will, to quote Burge (1986, p. 4), "help map the topography of a positive position", even if no clear positive position is advanced; just as Burge himself endorses no specific form of externalist theory of content (and endeavors to argue only that some form of externalism must be true), I will endorse no specific form of internalism.

To make my task more manageable, and to provide some organization to my analyses, I have tried to locate similarities in the strategies externalists have used; I then attack those strategies. I do not offer vague and general complaints about the broad drift of the strategies. What is interesting about these positions and arguments, and what is indicative of good philosophy generally, is the precision of the reasoning involved in them. My reactions to them, therefore, focus ultimately on the details of their execution. The positions and arguments I will be discussing are among the best examples of contemporary philosophy. The mistakes they make (if I am right in diagnosing them as mistakes) are often very small and subtle. But the beauty of philosophy is that small mistakes can bring down entire arguments. Philosophical arguments are like chains in that they are only as strong as their weakest link. However, whereas chains can be pulled and tested to see if they contain weak links, and if so where, philosophical arguments can be probed for weakness only upon laborious scrutiny. But therein lies the satisfaction of doing philosophy.

Part I

Externalism and the Propositional Attitudes

CHAPTER 1

EXTERNALISM AND TWIN EARTH

1. Preliminary Remarks

The dialectical situation, for my purposes, begins with externalist arguments, mainly by Tyler Burge, that purport to show that internalism is untenable. The arguments of central concern are those that begin with Twin-Earth-style thought experiments.[1] Anyone reading this has probably already heard of the Twin Earth thought experiments, and probably knows something about the arguments that the thought experiments have inspired. I will not try to produce here a comprehensive presentation and analysis of the literature that has grown up around the thought experiments. I will, however, say enough to make apparent the style of argument the thought experiments have inspired, and to convey the force of those arguments. Though I intend for this book to be approachable by those not already familiar with the literature, minimal familiarity with that literature is probably a prerequisite even for being interested in what I will have to say. At any rate, such familiarity will not hurt, and certainly is recommended, as the papers by Putnam and Burge on which the literature is founded are among the best examples of contemporary analytic philosophy.

2. Twins

In this section, I will lay out the Twin Earth style thought experiments, and the externalist arguments that have attended them. This will give the reader a sense of the substantial support for externalism, and provide a handy point of reference.

2.1 The Meaning of 'Meaning'

The style of the relevant externalist arguments is best introduced in the context of Putnam's (1975) discussion of externalism in semantics. Putnam was concerned with two standard claims about meaning:

(1) The meanings of words are fixed by the psychological states of those who use them.

(2) Meaning (or intension) determines reference (or extension).

The Twin Earth thought experiments were designed to help show that there is no meaning of "meaning" on which both (1) and (2) are true. In the thought experiment, there is a subject, S, on Earth who has an exact replica, Twin S, who resides on Twin Earth. Twin Earth is just like Earth, except that the liquid in the lakes, oceans, and rivers is not water (i.e., made of H_2O), but a superficially indistinguishable liquid twater (which is made of XYZ). S and Twin S know only the superficial qualities about the liquid in the lakes, rivers, and streams in the environment, and could not tell the difference between water and twater. Because Twin Earth is just like Earth in every other respect, when S requests a glass of water by saying "please bring me water", S's twin also utters the words (or, as Burge calls them, word forms) "please bring me water". However, whereas S's use of the term "water" means *water*, Twin S's use of the term means *twater*. That is, S has a concept *of* water, whereas Twin S has a concept *of* twater; the reference (or better, the extension) of their terms differ. By (2), which Putnam accepts, the meanings (or intensions) of S's and Twin S's uses of those terms must differ. Moreover, they differ in spite of the fact that their internal (and for Putnam, psychological) natures are identical.

Given that Putnam accepts (2), (1) must be false; the psychological states of the twins are not sufficient to determine the meanings of the words they use (because those psychological states are not sufficient to determine the references of those words). Putnam (1975, p. 223) expressed this conclusion by saying: "Cut the pie any way you like, meanings just ain't in the head". This has been the rallying cry of externalists ever since.

Putnam defended his externalist conclusion, at least as it applies to natural kind terms, by appealing to the fact that ordinary speakers of a language routinely defer to experts to fill out the meanings of terms with which they are otherwise competent. There is, as Putnam (1975, p. 228) calls it, a "division of linguistic labor" within any linguistic community. The important fact is that, in spite of not knowing the meanings of the terms we use, we still use them competently to refer to what the experts are themselves using the terms to refer to (see, e.g., Putnam, 1975, p. 246). Putnam surmised that we use the term "water", for example, to refer to anything that has the same essence as the liquid that is ostended in our "acquiring" the term. Putnam also argues that it is not open to us to claim that the meanings of the terms we use should be identified with the concepts, or "stereotypes", we associate with a term. Stereotypes change over time and from person to person, sometimes substantially, but it is implausible to suppose that the meanings of terms in a language are so variable. Moreover, as Donnellan (1966) and Kripke (1971) had pointed out, people can be refer to things without knowing all

or even much about the things they refer to. If we are to have objective identity conditions for meanings, we must bind them to the references, which do not change in the way that stereotypes often do (see, e.g., Putnam, 1975, p. 235).[2]

As a contribution to the semantics of public language, Putnam's externalism has a lot to recommend itself. But my interest is not primarily in semantics. Rather, I am concerned with the nature of mental states (e.g., concepts, propositional attitudes, and even the pre-conceptual contents of sub-doxastic and perceptual states). Burge, however, has extended Putnam's conclusion to the domain of propositional attitudes. I turn now to those arguments.

2.2 *Other Bodies*

Burge (1982) claims that the contents of propositional attitudes are given by the literal meanings of terms that occur obliquely in sentences attributing those attitudes. The term "water", for example, occurs obliquely in the content clause, or that-clause, of the following attribution.

S believes that water is wet.

Substitution of co-referring terms is not permitted in oblique contexts, nor is existential generalization.

Following on Putnam's semantic claims, the literal meaning of the term "water" in S's linguistic community differs from the literal meaning of the term in Twin S's community; *a fortiori*, the literal meaning of the term "water" at oblique occurrence in Earthly content ascriptions differs from the literal meaning of that term at oblique occurrence in Twin Earthly content ascriptions. Ordinary practice would attribute *water* beliefs to S using the word "water", but we would not ordinarily sanction the attribution of beliefs to Twin S that employ "water" at oblique occurrence; we would have to invent a term, "twater", that includes in its extension XYZ rather than H_2O, to use in attributions of the relevant thoughts to Twin S. S and Twin S would therefore have different concepts, and hence different beliefs, according to the common assumption about attitude attribution, in spite of their intrinsic, internal identity. As in Putnam's original case, the only variation is in the surrounding physical environment. Thus, not only does the meaning of one's use of a term depend on one's physical environment, one's very concepts and attitude contents depend on the environment in just the same way.

This claim is often cast in terms of the notion of supervenience. I will not enter into a detailed discussion of the supervenience relation. The main idea, which is sufficient for our purposes, is that a set of

properties P_S supervenes on a set of properties P_b if and only if there is no difference in P_S without a difference in P_b.[3] The claim to which Burge can be seen to object, then, is the claim that the property of having certain concepts and propositional attitude contents supervenes on states of the brain.[4] His argument appears to show, at least for propositional attitudes involving natural kind concepts, that the brain cannot form an adequate supervenience base; according to the thought experiment and Burge's psychological interpretation of it, we see a difference in concept despite no difference in intrinsic (i.e., brain) properties. If local supervenience is the right metaphysical relation between the property of having propositional attitude contents and the set of physical properties, then the smallest supervenience base for propositional attitudes involving natural kind concepts is the brain plus the local physical environment.

Burge supports his conclusion in part by appealing to fairly uncontroversial facts about concept acquisition and truth. Since S has been exposed to samples of water, while Twin S has been exposed to samples of twater, it is not surprising that S has acquired a concept of water rather than a concept of twater, and Twin S has acquired a concept of twater rather than a concept of water. There is no twater on Earth for S to form a concept of, and no water on Twin Earth for Twin S to form a concept of. And for this same reason, unless S has the concept of water, and Twin S has the concept of twater, a large number of their beliefs will be false. But such a conclusion is uncalled-for since both S's and Twin S's "beliefs...relate to their environments in exactly parallel and successful ways" (Burge, 1982, p. 110).

2.3 *Individualism and the Mental*

The structure of the argument we have just considered makes it apparent that the externalism for which it argues is not limited to natural kind concepts. In Burge (1979) we see that it applies with equal force to all sorts of concepts. Consider another Twin Earth example involving the concept of arthritis.

Suppose that, in the actual world, S believes that she has arthritis in her knee, and that it has spread to her thigh. She persists in this belief until, to her dismay, her doctor informs her that arthritis can only affect joints. This conforms to how "arthritis" is used in S's linguistic community. In a counterfactual Twin Earth world, Twin S has a belief that she would express as the belief that she has arthritis in her knee, and that it has spread to her thigh. She persists in her beliefs and, to her dismay, her doctor confirms her belief that the arthritis has spread to her thigh; in the counterfactual linguistic community, this is how the term "arthritis" is used.

In the actual case, of course, S's doctor is right about what "arthritis" means, and S is wrong. In the counterfactual case, though, Twin S is right about what "arthritis" means; it means something different from what it means in S's linguistic community. If the literal meaning of obliquely occurring terms in attitude attributions gives the content of the attitudes attributed, as Burge believes, then S's belief is a belief involving the concept of arthritis (as that is understood in the actual linguistic community), whereas Twin S's belief is a belief involving a different concept (since "arthritis" is understood differently in her community), the concept of, say, tharthritis (an English word coined to mean what "arthritis" means in Twin S's community). And again, this difference in concept and attitude cannot be attributed to anything internal to S or Twin S; it is only their environments, in this case their social (i.e., linguistic) environments, that differ.

Burge's main line of defense for his interpretation of this thought experiment focuses on the claim that ordinary practice would attribute to S a concept of arthritis that corresponds to the literal meaning of the term "arthritis" in spite of the fact that she has an incorrect understanding of the term.[5] He notes that ordinary practice favors attributions in which obliquely occurring words are taken literally. Following on Putnam's point, subjects do not protest that they have been misunderstood when corrected in the way that S has been corrected. Indeed, Burge (1979, p. 101) argues, idiosyncratic understandings that purport to capture the subject's ignorance often have no clear application. Moreover, he claims (1979, p. 101), it is not responsive to his arguments simply to say that the subject has an idiosyncratic understanding; the internalist must also show that the subject's words cannot or should not be interpreted literally.

2.4 *Intellectual Norms and the Foundations of Mind*

Burge (1986c) contains what is perhaps the most fully developed argument against internalism with respect to the individuation of intentional states. The argument involves the claim that meaning-giving normative characterizations of a term (synonymous expressions) are not indubitable. It is, in other words, false to say that if you understand a meaning-giving characterization, necessarily you know it to be true indubitably. The argument for this claim is complex and insightful in ways typical of Burge's writings. I will not discuss that argument here, however, because I propose to grant the conclusion. This may seem self-defeating, since Burge uses this conclusion as part of his argument that it is possible to possess the concept (or cognitive value) associated with the linguistic meaning of a term without understanding fully or correctly the meaning-giving characterization of that term, with which the term is synonymous. This secures the possibility that

one can be attributed a thought whose content is given in part by a word at oblique occurrence in a content clause, even though one does not understand fully or correctly the meaning of that term. This was the key step in the thought experiments and attending arguments of Burge (1979) and (1982).

To drive the point home, and extend its range of application, Burge offers another thought experiment that does not depend on a misunderstanding of the term (of the sorts so far considered), but on the endorsement of a deviant theory about the things in the extension of an ordinary general term. In this thought experiment, A in the actual world knows perfectly well that sofas are widely regarded as furnishings to be sat upon, so he has no delusions about the conventional linguistic meaning of the term "sofa". However, he comes to doubt the normative meaning-giving characterizations that articulate this view, and instead suspects that sofas are works of art or religious artifacts not at all suited for sitting. A encounters resistance when he expresses this suspicion. A has constructed an elaborate network of supporting beliefs and explanations in order to account for the *prima facie* evidence that does not bear favorably on his view (e.g., that people appear to sit on sofas and do not appear to revere them).

In the counterfactual world, A has a twin, B, who, like A, hears word forms that sound like the normative meaning-giving characterizations that cause A to believe that the conventional linguistic meaning of the term "sofa" involves the idea that sofas are pieces of furniture made or meant for sitting. In B's world, however, these word forms are not truisms of the sort that are considered to be meaning-giving characterizations; in fact, they are not true at all. In B's world, the objects that look like sofas are widely regarded to be works of art or religious artifacts; many of them would collapse if sat upon. But like A, and unlike most members of his own linguistic community, B believes that these sofa look-alikes (in English we can call them "safos") are widely taken to be pieces of furniture meant or made for sitting. And again like A, B begins to doubt what he takes to be the received view of these objects, and suspects that they are really works of art or religious artifacts. But unlike A, when B expresses his suspicion, he finds that people accept it as a matter of course.

So, A has correctly assessed usage in his linguistic community, but he has mistaken beliefs about sofas. B, on the other hand, has correct beliefs about safos, but he has mistakenly assessed usage in his linguistic community. According to Burge:

> 'The conclusion is that A and B are physically identical until the time when they express their views. But they have different mental states and events. A has numerous mental

> events involving the notion of sofa. B's skepticism does not involve thinking of anything as a sofa." (Burge, 1986c, p. 708.)

Thus, as with the previous thought experiments, intentional phenomena are seen to vary with variations in the external world, in spite of the lack of variation in internal constitution.

As with the previous thought experiments, the key point in Burge's interpretation is the claim that A has the concept that corresponds to the conventional meaning of the term "sofa" in spite of the fact that he doubts the meaning-giving characterizations of that term given in his linguistic community. This interpretation is supported by the fact that meaning-giving characterizations can be doubted coherently. In addition, this interpretation of the thought experiment is supported by the same considerations that support Burge's interpretation of the previous thought experiments. Each of them instantiate the same general form. First, they all assume that mental states are individuated by appeal to their contents. Second, in each thought experiment, the individuals' relations to their environments appear to be necessary for having the thoughts and concepts that they do. Third, the individuals' knowledge and discriminative capacities are not sufficient to allow the individuals to pick up on the environmental factors that figure in the individuation of their mental states (see Burge, 1986c, p. 709). There is, then, no reason why A cannot be attributed beliefs involving the concept of sofa. Similar considerations support the claim that B is thinking with a concept of safos. Once this is granted, externalism appears to follow.

2.5 Summary and Clarification

Burge has raised the issue of mental state individuation. Individuation is about this question: What facts make it true of an individual that she has a concept or propositional attitude of a given kind? The externalist answer is that facts about an individual's environment, more specifically facts about her history of relations to her environment, help fix the kind of state an individual has.

It is one of the assumptions of the twin-style arguments that kinds of concepts (and attitudes) are distinguished from each other by their contents; contents, then, are among the properties that individuate concepts and attitudes. The twin-style arguments also appeal to the fact that ordinarily we express the contents of concepts and attitudes by describing what they refer to. Those descriptions involve words that, in Burge's view, should be interpreted literally in attributing thoughts to ourselves and others; it is the literal meaning of those words in virtue of which they refer to the same things that our thoughts refer to, so it is the literal meanings of those words in virtue of which they express the

contents of thoughts. If there is indeed no way of expressing the contents of concepts and attitudes that is independent of their reference, then there is also no way of individuating concepts and attitudes that is independent of reference. The reference of such mental states is, then, essential to their identity. This is revealed to us when we consider individuals who are internally identical, but who inhabit different environments. Thus, the twin-style arguments appear to establish that the individuation of an individual's concepts and attitudes makes essential appeal to the individual's environment. In Burge's case, he has indicated more specifically that it is an individual's history of relations to the environment that is crucial to the question of mental state individuation; this follows from the fact that individuals learn concepts with empirical application by exposure to examples that are found in their environments.

3. *A Fantasy*

How might one object to the arguments for externalism? In this section, I would like to express a little fantasy, a glimpse at how an internalist must view the externalist arguments. This is a fantasy because the picture of the mind that it suggests is not yet theoretically complete. Much work remains to be done. Nevertheless, if it is my purpose to defend internalism, then I am obligated to give some sense of what it is that is being defended.

3.1 Semantics and Psychology

It has been a common assumption in the philosophy of mind that psychology needs semantics in order to deliver identity conditions for concepts and propositional contents (see Kobes, 1989, for an explicit claim of this sort). The idea is that the identity of, say, a concept is given by what the concept is *of*. Fodor (1993) and others have suggested, however, that psychology might not need semantics in order to provide identity conditions for concepts and propositional contents. In Fodor (1987) for example, this gets expressed as the claim that there is a kind of narrow content, which is a function from contexts to the anti-individualistic content that is given by literal interpretations of words at oblique occurrence in content clauses.

The problem with this notion of narrow content, however, is that it is notoriously difficult to express (see, e.g., Adams, 1993, for an explicit statement of this charge). It has been suggested (by, e.g., Fodor, 1982, p. 112) that we might use a description of what a subject takes, say, water to be (e.g., clear, potable, liquid), as a way of expressing the narrow content of a concept. But Adams (1992), Lepore and Loewer (1986) and others have suggested that such descriptions are no less anti-individualistic than the concept they are supposed to characterize. For

this reason, narrow content is usually taken to be content that is not about anything external to an individual. That is the only way such content can be individuated individualistically, and yet still have identity conditions; what it is about, and hence what gives it its identity, is not external to the individual. Thus, there is no violation of local supervenience. The anti-individualistic content, then, is content that does make reference to the external world; it is often called wide content, or broad content. As a result of the Twin Earth arguments, broad content is generally thought not to be locally supervenient.

Ideally, an internalist would like to be able to state the identity conditions for concepts in some way that does not get lost in the ineffable world of narrow content, but that does not violate local supervenience, either. What the internalist wants, in other words, is an account of broad content such that an individual has that broad content in virtue of properties that are internal to that individual; broad, locally supervenient content. In order to have such content, however, the internalist must find a way to express the content of a concept by some means other than an appeal to the literal meanings of words at oblique occurrence in content clauses (since that is, by the Twin Earth arguments, tantamount to accepting anti-individualism). By the same token, a simple-minded appeal to descriptions will not help either. Somehow, semantics and psychology have to come apart if internalism is to succeed.

3.2 *Against Literal Interpretations*

In saying that semantics and psychology must come apart, I mean to be claiming that, for internalism to succeed, there must be some way other than the appeal to the literal meaning of terms in psychological contexts to characterize the contents of thought. But why should we think that literal meanings are not appropriate means for characterizing the contents of thoughts?

Many have tried to justify repudiating Burge's commitment to the literal interpretation of words in psychological contexts. Bach (1987, 1988) and Crane (1991), for example, have claimed that Burge is ignoring a metalinguistic component in expressions of contents.[6] Loar (1988) makes a similar point in a nice way. He draws a distinction between social and psychological content. To see what he had in mind, consider Kripke's (1980) case of Pierre, who grew up monolingually in France, and believes that *Londres est jolie* (based on the testimony of others); upon learning English, Pierre comes to believe that London is not pretty (based on first-person exposure to the rougher parts of London). Since *"Londres est jolie"* translates into the English sentence "London is pretty", Pierre can be ascribed correctly the belief that London is pretty and the belief that London is not pretty. If literal interpretations of *d e*

dicto belief ascriptions are taken (as they are by externalists) to reveal the contents of Pierre's thoughts, then we would have to interpret the occurrence of "London" literally in the two belief ascriptions for purposes of ascribing the contents of Pierre's thoughts; Pierre, therefore, holds contradictory beliefs. On the assumption that Pierre is under no delusion regarding the contents of his thoughts, we must, quite implausibly, credit him with blatant irrationality.

Suppose that Pierre had visited, not a rough part of London, but a nice part, and formed the belief that London is pretty. Pierre therefore believes, for one reason, that London is pretty, and for another reason, that London is pretty. Loar asks, "...how many beliefs does Pierre have?" If *de dicto* ascriptions are to be interpreted literally for the purpose of ascribing content, Pierre would have but one belief, the belief that London is pretty. But obviously Pierre has here two beliefs, as evidenced by the fact that these beliefs would interact differently with other beliefs. Therefore, *de dicto* ascriptions should not, in general, be interpreted literally. Loar concludes:

> "These beliefs not only are individuated by commonsense psychology as distinct in their psychological roles; it also seems quite appropriate to regard them as distinct in *content*. The differences in their interactive properties flow from differences in how Pierre conceives things, in how he takes the world to be, in what he regards the facts as being -- that is, differences in some semantic or intentional dimension." (Loar, 1988, p. 103).

Similar remarks apply to the general terms that Burge has employed in his arguments against internalism. Burge's arguments depend on interpreting literally the terms that occur obliquely in the *de dicto* attributions in his thought experiments. As Loar has shown, however, there is, in general, no reason to interpret literally the obliquely occurring terms in *de dicto* ascriptions. Ordinary commonsense psychology demands otherwise.

Burge (1979, p. 93ff) resists such a position on the grounds that it requires implausibly that subjects have metalinguistic beliefs about the meanings of words that might differ from the literal meanings of the words. He claims that subjects should not be attributed such a metalinguistic beliefs. He seems to suggest that the literal meaning of a term is relevant to the individuation of the concept one associates with that term without the subject making any assumptions about the meaning of the term. He claims that there are clear cases where subjects do have metalinguistic beliefs of this sort, but the cases described in the thought experiments are not among them. The subjects in the thought experiments have beliefs about arthritis, water, sofas, etc., not about

the words "arthritis", "water", etc.; they have object-level beliefs, not metalinguistic beliefs. In other words, they do refer, not just in speech but also in thought, to the relevant things and stuffs; and they do so as effectively as those who are more competent with the terms. The individuation of their concepts is then constrained by this reference in that a difference in reference entails a difference in concepts. Thus, whereas the patient in the actual world has a concept of arthritis, the patient in the counterfactual world has a concept of tharthritis.

But, as Crane (1991, p. 18) points out, there is no accounting for how words and concepts get associated except in terms of beliefs or assumptions about how concepts should be expressed in one's language. The fact that these assumptions can be wrong testifies to their substance. Bach (1987, p. 266, fn.4) surmises, I think correctly, that if Burge talked, not about our metalinguistic beliefs, but about what certain terms mean to the subject, he might not have resisted the metalinguistic maneuver in the way he has. And as Bach also notes, Burge is not careful to distinguish between misunderstanding a notion or concept, and misunderstanding a term. The former, Bach maintains, is unintelligible, while the latter is not. At a minimum, Burge owes us an account of the relation between concepts and language that is devoid of metalinguistic commitments. In its absence it seems that Bach and Crane have a point.

One way to express the point that Bach and Crane and others have pressed is that there is a difference between speaker meaning and literal meaning. Thus, while I might refer to water by my use of "water", it may be that the concept I associate with that word would not, if used 'attributively', succeed in picking out water uniquely. If there were a way to characterize this sort of speaker meaning independently of the literal meaning of terms in psychological contexts, we might have a glimpse of a workable internalism.

3.3 Basic Concepts and Compositional Constructions

As we noted above, a simple-minded appeal to descriptions such as clear, potable, liquid will not help an internalist; like the literally interpreted words they are supposed to take the place of, these descriptions are, as many have noticed, anti-individualistic as well. But suppose our appeal to descriptions were a little more sophisticated....

The descriptions in question are intended to be analyses of the individual's concept of, say, water. If S thinks of water as a clear, potable, liquid, then that would be a decent first approximation of the content of S's water concept. Suppose, now, that instead of resting with that description, we provide analyses of how S thinks of clarity, potability, and liquidity. And suppose that we provide analyses of these further descriptions, iteratively, all the while characterizing

how S thinks of the concepts involved in the analysis. And now suppose that these iterated analyses eventually dry up by reaching a subset of a larger set of innate, primitive concepts. And suppose that these innate, primitive concepts are concepts that have their content solely in virtue of the internal physical properties of our brains. We may not understand the way in which physical properties give rise to intentional properties such as this, but, then, no one in this debate has any idea how anything can have intentional properties (so this is hardly a strike against the story being told here). The fact just may be that we are blessed with such innate, primitive concepts, and that these concepts are *about* the external world. That is, these basic concepts are not narrow in the sense of referring to nothing beyond the individual's skin; they do refer to real properties of the external world, so they are broad, not narrow. But, because they have their broad content solely in virtue of factors internal to the individual, their broad content is individualistic, not anti-individualistic.

Given that these are the concepts that figure in the analyses of our more complex concepts, we can suppose that these more complex, less primitive concepts are compositional constructions out of the more basic concepts.[7] And, because the basic concepts are broad, the compositional constructions out of these concepts are also broad. But, just as the basic concepts are individualistic despite being broad, the complex concepts are both broad and individualistic as well.

Note that none of the usual difficulties for a descriptive theory of reference (and hence the arguments for a direct theory of reference) apply here because this picture does not call into question the literal meanings of the words (see, e.g., Salmon, 1981, for a nice summary of the arguments for the direct theory of reference). I do refer to water whether I know its defining features or not; the semantics of my words remains unaltered.

My psychosemantics, and therefore, also, the identity of my *concept*, is quite another matter. The content of my concept of water is to be given by a description that eventually bottoms out in innate, basic concepts.[8] If I have learned my concept of water well, then, at least in this world, it will be about water; that is, the set of descriptions that constitutes my concept of water will, when used attributively, pick out water on Earth. My twin will have the same set of descriptions, and hence the same content for his concept of twater, except that on his planet, the descriptions pick out twater rather than water. This is an interesting semantic difference, but not one that affects the psychological identity of our concepts since our set of descriptions is the same. And, again, since each is the compositional product of basic concepts whose content is determined solely by internal features, our

concepts, which are identical in content, are also determined solely by internal features.

4. Concluding Remarks

This, then, is my fantasy. It is a fantasy of broad content that is individuated wholly individualistically. It supervenes solely on internal features, it is determined solely by internal features, it is the content it is solely in virtue of internal features; it is individualistic by any measure. And still it is broad content. But this *is* a fantasy. There are many components in the story that have not been worked out. Though there are some reasons for thinking that words that occur obliquely in content-clauses should not necessarily be interpreted literally, those reasons are controversial. Moreover, while I have argued for the compositionality of mental content, those arguments have yet to win adoring approval from the philosophical community. Worse still, *no* reason has been given for thinking that there are innate concepts whose broad intentional properties are determined (metaphysically) by internal physical properties of brains; and no one has even the slightest idea how such an emergent property could exist.

My goal throughout the remainder of this book, however, is to show that there is no reason not to buy into this fantasy. The externalist arguments we have already considered are silent here since it is a key assumption of them that the internalist is questioning. All other arguments for externalism, I will argue, fail to burst the bubble of this fantasy. In the next two chapters, I will argue for a more aggressive position. I intend to show that externalism is burdened by quite implausible epistemological implications; it is inconsistent with widely accepted views of self-knowledge, and it appears to entail that introspection is sufficient for knowledge of empirical claims (in which case skepticism would turn out to be, not just false, but incoherent). It is the implausibility of these implications that, I suggest, casts doubtupon externalism about the propositional attitudes. If externalism is indeed false, then there must be something wrong with the twin-style arguments for it. And, while I would dearly love to articulate a positive internalist thesis in more detail than the fantasy just expressed, and expose just what precisely is wrong with the twin-style arguments, that is book-length project in its own right (a project, however, that is already under way). The near-term goal is simply to make the need for that future project more acute.

CHAPTER 2

SELF-KNOWLEDGE

1. Preliminary Remarks

Freudian worries to one side, it is fair to say that we know a lot about ourselves. I know who I am, what type of person I have become, what I have made of myself in life, and even, to some extent, what I would like to do with the rest of my life. And even if I had never thought about such lofty matters, all I would have to do to acquire knowledge of such things is reflect for a time on my thoughts, memories, and aspirations; all I would have to do is think. I understand my motivations, at least some of the time, even when they are less pure than I would like them to be. I do not have to ask anyone else what it is that I think about the state of affairs in the world today (though I am often tempted to defer to others in determining what I *should* think). In a wide range of respects, we all take ourselves to be the authority, if not on any other subject, at least on the subject of ourselves.

This self-knowledge is, in the first instance, rooted in our knowledge of the contents of our thoughts; we could not know what we know about ourselves without knowing what it is that we think when we think it and reflect on it. There is a peculiar authority with which we know the contents of at least some of our thoughts. Though there is some obscurity in the idea that we know our thought contents in this way, it seems undeniable that, under certain circumstances, we know the contents of our thoughts non-inferentially, without having to perform any empirical investigations of the external world. The claim of self-knowledge is not the claim that our knowledge of the contents of our thoughts is infallible, nor is it the claim that self-knowledge is complete, nor even is it the claim that we have such self-knowledge all the time. Rather, it is the fairly humble claim that, when we do know the contents of our thoughts, we know those contents in a direct and privileged way, without inference or empirical investigation. It is, in other words, the claim that, for a wide range of empirical thoughts (thoughts about the external world), we are not *systematically* debarred from knowing what it is that we are thinking.

A significant amount of recent philosophical attention, however, has been directed at the relationship between externalism and self-knowledge.[1] The worry is that externalism will turn out to be

inconsistent with even these limited, but plausible, claims about self-knowledge; that is, if externalism were indeed true, there would seem to be indefinitely many empirical thoughts whose contents we would not be in a position to know in any privileged way. Many externalists, e.g., Falvey and Owens (F&O), Burge, Heil, and others, have conceded that if such a case can be sustained, it will show that externalism must be mistaken. There is a "puzzle", as Burge puts it, involving externalism and self-knowledge, one the externalist must dispel.[2] I argue in this chapter that externalism is indeed, to its detriment, inconsistent with authoritative self-knowledge. I will also try to show, contrary to some (Boghossian and Heil), that internalism faces no corresponding difficulty.

Much of what I have to say about the nature of self-knowledge will emerge in subsequent sections. In order to ward off possible misconstruals and improve the clarity of the discussion to follow, however, it would be prudent to begin with a few preliminary remarks about the sort of self-knowledge at stake in this dispute.

To start, I want to distinguish the sort of claim I defend from Russell's claim that thinking a particular thought *requires* knowledge of what the thought is about (see Russell, 1912, p. 58). Russell's principle, in the first instance, seems to concern knowledge of the referents of thoughts, what the thoughts are about, not knowledge of the contents or meanings of thoughts, those aspects of thought that determine reference. If this is indeed so, it might then be argued that the principle is simply not true; we do not need to *know* anything about the referents of our thoughts in order to have the thoughts in the first place; mere belief is sufficient. But, of course, if contents are individuated by appeal to what they are about, then Russell's principle may well be, in some sense, true. Nevertheless, I will not be concerned with Russell's claim. The real reason I will not be engaging it is that it is about the very conditions necessary for thought; having the thought that x is F requires one to be acquainted with x and F-ness. I do not take the sort of knowledge of content at issue in the following dispute to be a necessary condition on having the thoughts in the first place. It is, I believe, a contingent fact about us that we do know, to varying degrees, the contents of some of our thoughts. It is possible to have these same thoughts without enjoying any second-order thoughts about them. So, insofar as Russell's principle is a principle about the necessary conditions on thoughts, it does not bear on the much weaker claim that I will be discussing.[3]

I will deploy the familiar distinction between two components of propositional attitudes, the propositional *content* (often identified by expressions in a language, such as, e.g., water is wet) and the *attitude* we take toward that content (e.g., belief). Our concern in this chapter is not

with our knowledge of the attitudes we take toward contents, but with our knowledge of the contents of our thoughts.[4] I do not take a stand on, or have a stake in claiming anything about, how well we know the attitudes we take toward particular contents.[5] I can imagine, though, a case where I fear that I am about to get hurt by an on-rushing bull, but, due to some conceptual poverty, I do not know that I am in a state of fear. I do not, however, think that skeptical claims about my knowledge of the contents of that fear-thought are nearly as plausible. At any rate, since psychosemantic externalism is a thesis about the contents of our thoughts, it bears only on skepticism about knowledge of contents, not skepticism about knowledge of attitudes. There might be good reasons to be skeptics about our knowledge of the attitudes we take toward particular contents, but it is hard to see how they could involve externalism.

More specifically, my concern in this chapter is with *occurrent* thought contents, not the sort of dispositional thoughts that seduced behaviorists. The claim that we have a privileged sort of knowledge of the contents of our occurrent thoughts famously goes back at least to Descartes, who made many bold, and implausibly strong, claims about our access to the contents of our thoughts (see, e.g., Heil, 1992, pp. 152ff for discussion). Whatever his excesses, Descartes was surely not wrong to think that we know the contents of our thoughts in a direct sense in which others do not know the contents of our thoughts. Moreover, we know the contents of our thoughts in a direct sense in which we do not know the nature of the external world; there is, as many have noticed, an epistemic asymmetry in our knowledge about our own minds and our knowledge of the external world.

The directness with which we know the contents of our thoughts is an *epistemic*, and not merely *causal*, directness. We do not have to make inferences to determine what we are thinking; we can know what we are thinking, as McKinsey (1991, p. 16) puts it, "just by thinking". This is not to suggest that we cannot draw inferences about what we are thinking; in some cases, those that perhaps interested Freud, we can discover something about the contents of our thoughts, and the attitudes we take toward them. I will therefore remain neutral on questions involving repression and self-deception. I will *not* remain neutral, however, on the claim that there are large ranges of occurrent thoughts whose contents are not in any way concealed from us. The question will be whether externalism (or even internalism) implausibly requires that we would have to make inferences or empirical investigations of the external world in order to know the contents of those occurrent thoughts. Should externalism imply that we can know our thoughts only in this indirect way, it will imply a falsehood, and thus be open to refutation by *reductio* (as, e.g., F&O, 1994, p. 108, concede).

2. *Externalism and Knowledge of Content*

Boghossian (1989) argues that there is *prima facie* reason to take seriously the claim that there is a conflict between externalism and the purported directness and authority with which we know the contents of our thoughts. In this section, I will concentrate on defending the sort of argument Boghossian advances. The claims involved in this argument raise deep and complex issues at the intersection of semantics, epistemology, and the philosophy of mind that have not, thus far, received adequate treatment in the literature.

2.1 *Opening Arguments*

Here is a simple version of the argument: If the very contents of the thoughts we think depend for their identity on facts about the physical and/or social environment as externalism maintains, then any knowledge of them (knowledge of what thoughts they *are*) would seem to require knowledge of the environments on which they depend. How, for example, could we determine whether we are thinking the thought that *water* is wet rather than the thought that *twater* is wet without knowing something about the environment, namely, that it contains water and not twater? If externalism is right, it would appear that we have to know whether our world contains water rather than twater if we are to know what content our thoughts have. But this, of course, would be empirical knowledge, and is therefore not introspective, and neither direct nor authoritative. So, externalism requires us to reject claims of direct authoritative self-knowledge.

Burge (1988b, p. 654-5) recognizes, and Boghossian (1989, p. 12) acknowledges, that this argument moves too quickly. It assumes a characteristically Cartesian epistemological internalism, according to which knowledge that p requires that we be able to rule out any chance of error; if there are any other conflicting claims that might (for all we know) be true, we cannot be said to know that p. Epistemological externalists (e.g., Goldman, 1976) challenge this claim by weakening the conditions under which one can be said to know that p. According to epistemology in this tradition, in order to have empirical (e.g., perceptual) knowledge that p, we are expected to be able to exclude only "relevant" alternatives to p. For example, we would not have knowledge that there is a glass of water on the bar of the saloon if we cannot exclude the possibility that the glass is full of gin; the gin-hypothesis is a relevant alternative in the context of a saloon (even if, say, it would not be a relevant alternative in the context of, say, an athletic locker room). However, we would be entitled to claim knowledge that there is a glass of water on the bar even if we cannot

exclude the possibility that the glass is full of twater; the twater hypothesis is not a relevant alternative.

Now apply this reasoning to the case of knowledge of content. It would seem to be true that in order to know that we are thinking that water is wet we must be able to tell the difference between the thought that water is wet and the thought that gin is wet; the gin thought is a relevant alternative. We must admit, however, that we can know that we are thinking that water is wet even if we *cannot* tell the difference between the thought that water is wet and the thought that twater is wet; twater thoughts just are not relevant alternatives.

Boghossian (1989, p. 13) is quick to point out, and Burge (1988b, p. 652) and F&O (1994, p. 115) recognize, that it is easy to contrive cases where the twater-thought becomes a relevant alternative. So-called slow-switching cases, where an individual is transported surreptitiously between Earth and Twin Earth, are cases where one can legitimately wonder whether one is thinking the thought that water is wet rather than the thought that twater is wet. In these cases, externalism seems clearly to preclude knowledge of content. Thus, Boghossian's argument appears to have force even if one is working with an externalist epistemology in the relevant alternatives tradition; to the extent that claims about knowledge of content in these cases are secure, externalism is not.

But the argument does not stop here. The general claim that externalism is in conflict with home truths about self-knowledge has been contested along many fronts. Given the controversy that has been stirred up by the sort of argument Boghossian advances, one might begin to wonder whether externalism really does face the unwelcome implication that we lack knowledge of the contents of our thoughts. To restore a firm sense of the difficulties facing externalism on this issue, there is no other way to proceed than to show that prominent externalist apologies are uncompelling.

2.2 *First Defense of Externalism: Warfield on Relevant Alternatives*

I begin the rehabilitation of the Boghossian argument by looking at Warfield's (1992) attempt to defeat it. Warfield argues that Boghossian's suggestion for circumventing the appeal to relevant alternatives is flawed. Warfield acknowledges that "...relevance is an objective notion and an alternative's being actual is sufficient for the alternative's being relevant." (Warfield, 1992, p. 234) It is true, then, that the twater thought is a relevant alternative in the switching case Boghossian describes. Nonetheless, he claims, Boghossian's argument is beset with difficulties. Warfield interprets the argument in this way:

P1 To know that P by introspection, S must be able to introspectively discriminate P from all relevant alternatives of P.
P2 S cannot introspectively discriminate water thoughts from twater thoughts.
P3 If the Switching Case is actual, then twater thoughts are relevant alternatives of water thoughts.
C1 S doesn't know that P by introspection. (Warfield, 1992, p. 234-5)

In Warfield's view, the argument is invalid, and establishes only

C1' If the Switching Case is actual, then S doesn't know that P by introspection. (cf. Warfield, 1992, p. 235)

But C1' does not bear on the compatibility of externalism and introspective self-knowledge, and instead bears "*at most*" on this question:

Q Given externalism, is it *necessary* that the contents of a thinker's thoughts are knowable to the thinker on the basis of introspection? (cf. Warfield, 1992, p. 235)

Thus, Warfield concludes, Boghossian has done nothing to circumvent the relevant alternatives approach, and save the argument for the incompatibility of externalism and introspective self-knowledge.

Warfield, however, fails to make his case. He notes explicitly (p. 234) that the actuality of an alternative is *sufficient* for its being relevant. But the fact that being actual is *sufficient* for relevance does not imply that it is *necessary*. Specifically, twater thoughts do not have to be actual to be relevant, even though, were they actual, they would of course be relevant; twater thoughts may be relevant even if they are only possible thoughts S might have. Without giving us a deeper understanding of the relevance relation, Warfield has no justification to limit relevance to thoughts *actually* thought, while ruling out thoughts one *might* think. Therefore, Warfield has no warrant to canvass Boghossian's argument with P3 as opposed to P3':

P3' Even if the Switching Case is not actual, twater thoughts are relevant alternatives to water thoughts.

And if P3' is true (and more faithful to Boghossian's argument), C1 follows. This is not yet to say that P3' *is* true; the *mere* possibility of twater thoughts does not guarantee their relevance. But because

Warfield has not shown that P3' is false (or even implausible), he has not shown that Boghossian's argument is unsuccessful. Should support for P3', or an appropriately general analogue to it, be located, then Boghossian's argument, as he intended it, not as Warfield has canvassed it, will go through.

Moreover, even if we were to grant that C1' can be concluded from the argument, it is not limited to bearing only on Q, so long as it is also true that

> C1" Even if the Switching Case is not actual, then S doesn't know that P by introspection.

C1' and C1" together bear on the compatibility of externalism and introspective self-knowledge; they plainly entail incompatibility. It is plausible to suppose that Boghossian had in mind something like this. The appeal to switching cases in his argument might simply be to illustrate that there are circumstances where twin thoughts (not just water and twater thoughts but twin thoughts in general) are relevant alternatives. Perhaps all he intended to claim is that if there is a realistic (as opposed to mere) possibility of a switching case arising, that is sufficient to show that those twin thoughts are relevant alternatives (I will have more to say about this line of reasoning below). Since Warfield has given no argument against C1", he has failed to demonstrate that the conclusion of the argument is limited to bearing at most on Q. He has therefore not sustained his objection to Boghossian's argument, and has not justified the claim with which he concludes his paper, i.e., "Privileged self-knowledge and externalism are compatible."

Of course, if the conclusion had instead been

> C1"' *Only* if the Switching Case is actual, S does not know that P by introspection,

then Warfield would be correct in limiting the bearing of the conclusion to Q. But, for all Warfield has shown, twater thoughts are relevant without the switching case being actual; therefore, for all Warfield has shown, externalism entails that S does not know that P by introspection (regardless of whether the switching case is actual). For all Warfield has shown, then, the truth of externalism is incompatible with the truth of the self-knowledge claim.

But this point is cold comfort to the opponent of externalism unless it can also be shown how switching cases can bear on the issue of the compatibility of externalism and introspective self-knowledge without being actual. We can take a first step in this direction by

considering Warfield's question Q, which asks whether introspective knowledge of content is necessary. I contend, in response, that even if the most that could be established is a negative answer to Q, as far as externalism is concerned *a negative answer to Q is bad enough.* For if one were to answer Q in the negative (as Warfield's remarks suggest he is inclined to do), then one would be conceding that sometimes an individual might not know the contents of her thoughts by introspection simply because knowledge of those contents requires investigation of the environment. I am prepared to grant that introspective knowledge of content is not infallible.[6] But a negative answer to Q entails more than this. It entails that introspective self-knowledge is not infallible *because* knowledge of content sometimes requires investigation of the environment. However, as Warfield himself points out:

> "What motivates philosophers' worries about self-knowledge and externalism is the intuition that externalism might imply the seemingly false conclusion that one must investigate one's environment to know the contents of one's thoughts." (Warfield, 1992, p. 236)

A negative answer to Q admits that there are indeed cases where knowledge of content requires precisely that. But if it is "seemingly false" to suppose that introspective knowledge of content in general should require empirical investigation of the environment, then it would seem just as false that we should *ever* have to investigate the environment to know what we think. If there are any plausible circumstances under which externalism entails that we have to investigate the environment to know what we think, then externalism would seem to be just as a ripe a target for a *reductio* as if it entailed that we must always investigate the environment to know what we think. Introspective self-knowledge may be fallible,[7] but it should not be fallible for *this* reason.

One might wonder at this point whether there ever *are* any plausible circumstances under which externalism would have these unwanted epistemological implications. That there are such circumstances can be seen by noticing that switching cases are not mere abstract possibilities; sometimes they are indeed actual. And because they actually do tend to occur, they are often possibilities that must be considered, so the alternatives to which they give rise are often *relevant* alternatives. This is parallel to questions of ordinary empirical knowledge, where possible types of circumstances that actually do tend to arise make the alternative possibilities they generate relevant; e.g., now that terrorists actually do strike in U.S. cities, the possibility of a terrorist strike, and the alternative states of

affairs to which such a possibility give rise, is relevant to questions about our knowledge that we are safe in government buildings.

Ludlow (1995) has argued that for social externalism of the sort that Burge (1979) defends, switching cases do not occur only rarely, but instead are prevalent. According to Burge's social externalism, the meanings of our words, and the contents of the thoughts we express with those words, depend on the literal meanings of the words in our language community; we often defer to others when our knowledge of the meaning of terms is incomplete. Switching cases that are relevant to social externalism arise when one "unknowingly slide[s] from one language community to another". This, Ludlow demonstrates, is common as we travel between different cultures, or even as we move from one circle of acquaintances to another. Take, for example, a native speaker of American English, S, who travels to Britain and stays for an extended period of time; suppose she takes "chicory" to mean the same in British English as it does in American English, though she has only incomplete knowledge of the meaning of the term; according to Burge and Putnam, she then defers to her language community, particularly the 'experts', who fix the meaning of the term. As a matter of fact, however, "chicory" means something different in American and British English. Because S's language community has become relevantly different, S is a victim of slow switching (cf. Ludlow, 1995, p. 47).

According to social externalism, then, when S defers to her (now British) language community to fix the content of thoughts she would express with "chicory", she ends up thinking something other than what she thought when her language community consisted of Americans. Nevertheless, because she is a victim of slow switching, her American chicory thoughts are of course *relevant* alternatives to her new British chicory thoughts. And yet, because the difference between the thoughts lies in her language community, she cannot discriminate between the thoughts, and hence cannot know that she is thinking one thought rather than the other, without investigating the practices of her language community; even by the relevant alternatives approach, she does not know the contents of her thoughts by introspection alone.[8]

This point is very similar to one made by Davidson (1987, e.g., p. 448), who maintains that if S is wrong or confused about the meaning of the words she uses (or would use) to express her thoughts, owing to their essentially social identity conditions, then Burge is committed to the concession that she is wrong or confused about (and hence does not know) what she believes. Burge thinks Davidson is wrong, and would presumably have the same reaction to Ludlow's point. Burge (1988b, p. 662, also fn. 10) diagnoses Davidson's error to be participation in the view that knowledge of content requires one to be able to explicate the contents of one's thoughts. But regardless of whether knowledge of

content requires or involves explicational abilities, Ludlow's development of Boghossian's discussion of relevant alternatives serves to show that Davidson's point does not rest on assimilating self-knowledge to explication; it obviates Burge's response by showing how S's "partial understanding" of her thought contents undermines her *knowledge* of thought contents even on an externalist epistemology of the sort that Burge himself favors (see, e.g., Burge, 1988b, p. 655, esp. fn. 6).

Ludlow goes on to describe several even more common situations in which a subject may be victim to cases of slow switching between language communities. According to Burge's more recent writings, however, our thoughts are not just dependent on our linguistic environments, but also on our physical environment; we here on Earth have the concepts of water and aluminum, rather than twater and twalum, because we have been exposed to water and aluminum in learning our concepts (see, e.g., Burge, 1982, 1986, 1988b, 1989). It would extend the range of our discussion if we consider the prevalence of slow switching cases between appropriately different physical environments.

Assume (what I believe is true) that there are subtle differences in constitution, too small really for non-expert cola drinkers to notice, between the colas sold by a certain brand X in Europe, and the colas sold by that same brand X in the U.S.; there are, in other words, E-colas and U-colas. Now imagine that S is an average cola drinker who has spent a fair amount of time in both Europe and the U.S., and has all the while been ignorant of the systematic differences between E-colas and U-colas. Externalism seems to entail that S will have developed, unwittingly, two concepts, one of E-colas and one of U-colas. But suppose that S thinks the thought she would express as "colas taste good". What concept, according to externalism, is deployed in this thought? Regardless of which concept is in fact featured, it is clear that the other would be a relevant alternative. So, if S cannot tell the difference between E-cola-thoughts and U-cola-thoughts, S cannot be said to know the contents of her thoughts; such is the dictate of externalism.

This example is significant if only because it has application to a large number of people who have sampled colas in the U.S. and Europe; for even this shows that externalism entails that enormous numbers of us sometimes fail to have introspective knowledge of the contents of some of our thoughts due simply to their environmental determination. More importantly, however, this particular example instantiates a type of example that is extremely widespread. Types of physical objects and substances, whether natural or artifactual, display a hierarchical structure; many types come in sub-varieties that bear superficial similarities to each other, but are nonetheless of different lower-order types (with, e.g., different internal constitutions). Many of

usare such that we cannot make out differences between types of beers, frozen pizzas, cheeses, gravel, grass, flowers, sand, and indefinitely many other types. Whenever, e.g., varieties of beer are clustered locally such that one learns one's concept of beer initially by exposure to one variety, but later interacts long enough with another (and fails to be aware that such a difference exists), a slow switching case is likely to arise.

An externalist may try here to deny that all such differences in types of physical objects beget differences in types of concept. But such a response requires that the externalist supply much more careful principles of concept and content individuation than have so far been offered. When is the underlying nature of the physical environment relevant to concept individuation and when is it not? Surely not every minute physical difference is relevant, but externalists have so far provided no clues as to how they propose to distinguish relevant differences from those that are not relevant (except, perhaps, in the context of natural kinds). Nor is it clear that such principles are forthcoming. Why, for example, should natural kind differences matter, but not differences between colas? But even if there are plausible such principles on the horizon, and especially if there are not, the externalist must admit that slow switching cases are likely to be prevalent since no principles will eliminate the many respects in which hidden physical differences make for differences in concepts. That is, externalism *must* admit that such cases can arise since the very nature of Twin Earth thought experiments requires them. A more detailed account of when such environmental differences make for differences in concept will only expose more clearly just which cases are the sorts of slow switching cases that pose a problem for externalism.

The point of discussing these cases is that their ubiquity makes them more than just abstract possibilities. They can and do occur (and the externalist must admit that they occur), so the alternative thoughts to which they give rise are often *relevant* alternatives, even if in a particular case no slow switch has taken place. This is parallel to the issue as it arises in the context of empirical knowledge; circumstances that tend to occur give rise to relevant alternatives, even if they do not actually occur in a particular instance, simply because their general prevalence indicates that they might easily have occurred. Thus, we can interpret Boghossian's argument according to P3', rather than Warfield's P3, and defend that premise, as Ludlow does, by pointing out the prevalence of actual switching cases. Thus, there is reason to think that both C1' and C1" are true. If we accept Boghossian's appeal to relevant alternative epistemology, externalism entails that frequently we lack introspective knowledge of the contents of our thoughts; for according to externalism, these alternative thoughts cannot be

discriminated introspectively from the thoughts one actually thinks. Therefore, even according to epistemology in the relevant alternatives tradition, externalism is incompatible with introspective self-knowledge.

2.3 Second Defense of Externalism: The Standard Strategy

Burge (1988b), however, has anticipated a move like this, and he has given an interesting response to it. I will call this 'the standard strategy' since it is, in part, rather similar to strategies adopted by several others.[9] The standard strategy has two main components. First, it involves the idea that the second-order thoughts that are candidates for self-knowledge inherit the contents of the first-order thoughts that they are about. Having met the conditions for thinking the first-order thought that water is wet, one has thereby met at least some of the conditions for thinking the second-order thought that one is thinking that water is wet. The only further conditions required are those required for the generation of second-order thoughts at all; there is no further condition on getting a second-order thought with the 'right' content. As Burge (1988b, p. 654) puts it, "[a] knowledgeable judgment that one is thinking that water is a liquid must be grounded in an ability to think that water is a liquid." If one is situated with respect to one's physical or social environment such that one forms the first-order thought that water is wet, then if one also has the ability to think second-order thoughts, one will thereby be situated appropriately so that the second-order thought will involve the same concepts as the first-order thought; no investigation of the environment is required to ensure that one is thinking a thought with the externally individuated concept of water as opposed to the concept of twater. The content of the first-order thought is inherited by the second-order thought even though no steps are taken to verify the identity of the concepts involved (e.g., by examination of the environment).

Burge takes this position a step further by focusing more narrowly on a pure form of self-knowledge that is entirely immune to error. According to this refined position, in order to know that one is thinking that water is wet, one must think the thought that water is wet in a certain "reflexive" way. When S thinks the second-order thought that she is thinking that water is wet, she thinks the first-order thought and attributes it to herself. She cannot mistakenly think, in her second-order thought, that she is thinking some other first-order thought because the very thinking of the second-order thought involves the thinking of the first-order thought; she thinks that water is wet "...in the very event of thinking knowledgeably that [she] is thinking it" (1988b, p. 654). Burge calls such second-order judgments that are not numerically distinct from the first-order thought "basic self-

knowledge". These reflexive judgments, Burge maintains (e.g., 1988b, p. 658), are "self-verifying", and the impossibility of error in these judgments bestows upon them their peculiar authority, directness, and privileged status; no comparisons to alternative thoughts, even relevant alternative thoughts, could undermine their authority, and no empirical investigation of the environment is required to establish their veracity.[10]

Thus, knowledge of the contents of one's thoughts does not entail that one can discriminate those contents from all possible contents, most especially twin contents. The second component of the standard strategy, then, is to distinguish between introspective knowledge of content (KC), which involves knowing directly the contents of one's thoughts, and introspective comparative knowledge of content (CKC), which involves knowing of two or more thoughts whether their contents are the same or different.[11] It is obvious that externalism is incompatible with introspective comparative knowledge of content. If thought contents are determined in part by environmental factors, then the twins in the Twin Earth thought experiments will have different thoughts; but, as Burge (1988b, p. 653) readily admits (and even insists), their thoughts will be introspectively indiscernible. Therefore, if externalism is true, we will not be able to tell by introspection whether or not the two thoughts have different contents. But, in part because comparative knowledge of content is a different matter from knowledge of content (i.e., CKC is not the same thing as KC), it does not follow from the fact that externalism implies that we lack comparative knowledge of content that it implies that we also lack knowledge of content. Given this distinction, it is perfectly possible that, even if externalism is true, we can have knowledge of the contents of our thoughts.

The rejection of CKC as a condition on KC is motivated in part by the foregoing analysis of KC, in which the first-order thought is a component in the second-order thought; knowledge of content, insofar as it requires the impossibility of error, does not require ruling out alternatives of any kind. The rejection of CKC as a condition on KC is also motivated by a limited comparison to ordinary perceptual knowledge. Externalist theories of knowledge, such as the relevant alternatives approach, eschew the requirement that one must be able to rule out the possibility that one is mistaken; one may know that p even though one can only rule out a limited range of conflicting possibilities. Standard strategists argue that the requirements for self-knowledge should be no stronger than this. Thus, it may be that KC is, in general, a kind of self-knowledge that we enjoy, and it may also be compatible with externalism, even if CKC is, in general, not. Boghossian's argument, therefore, cannot be successful.

2.4 An Unsuccessful Internalist Rejoinder: Content Skepticism

In a recent reply to F&O, Brueckner (1994) claims that what I am calling the standard strategy does not take account of *how* the relevance of twin thoughts shows that there is a conflict between externalism and KC. If correct, Brueckner's points would undermine the standard strategy and save Boghossian's argument that there is a conflict between externalism and KC. But Brueckner's points are not correct; in showing why they are not correct, we can get some sense of the tenacity of the standard strategy.

F&O endeavor to extract the principle that underwrites Boghossian's treatment of the appeal to relevant alternatives. They first suggest the following principle.

> (RA) If (i) q is a relevant alternative to p, and
> (ii) S's belief that p is based on evidence that is compatible with its being the case that q, then
> S does not know that p. (From F&O, 1994, p. 116)

A reconstruction of Boghossian's argument based on RA would then look like this:

> (a) Externalism is true
> (b) The thought that twater is wet is a relevant alternative to the thought that water is wet (due to consideration of the switching cases)
> (c) Given externalism, S's belief that she is thinking that water is wet is based on evidence that is compatible with its being the case that twater is wet
> (d) Therefore, S does not know that she is thinking that water is wet.

This argument appears to be valid. But F&O believe that RA itself is underwritten by RA', which is identical to RA, except that (ii) becomes (ii'):

> (ii') S's justification for her belief that p is such that, if q were true, then S would still believe that p.

We must, therefore, reconstruct the argument again. In the new version, (c) becomes (c'):

> (c') S's justification for her belief that she is thinking that water is wet is such that, if she were thinking that

> twater is wet, then she would still believe that she is thinking that water is wet.

But now, F&O point out, the argument is no longer sound; (c') is false, and without it the most that could be established is a rejection of CKC. S does not know that she is not thinking the thought that twater is wet; therefore, she does not know that she *is* thinking the thought that water is wet *as opposed to* the thought that twater is wet. But this is a failure of CKC, not KC. Thus, the distinction between KC and CKC allows the externalist to maintain that, even accepting that Boghossian is right that twater thoughts are relevant alternatives to water thoughts, the argument does not threaten KC. And since KC is not threatened, there is no further *reductio*-style threat to externalism.

Brueckner labors under the impression that F&O have assimilated the reasoning from RA' to Brueckner's own similar argument in his (1990). He then proceeds to claim that he did not reason in this way, and that to suppose that a "content skeptic" would reason in this way is to be uncharitable to the content skeptic. His contention, therefore, is that the standard strategy's appeal to the distinction between KC and CKC attacks a straw argument.

To demonstrate this, Brueckner (1994) reconstructs an argument along the lines of RA', on which (c') is replaced by something like:

> (c'') S's justification for her belief that she is not thinking that twater is wet is such that, if S were thinking that twater is wet, then S would still believe that she is not thinking that twater is wet.

Brueckner then realizes that this only supports a subconclusion:

> (d') S does not know that she is not thinking that twater is wet.

Thus, to draw conclusion (d), he says, one must argue from (d') by means of the following principle:

> Closure: Knowing that p requires knowledge of those consequences of p that one knows to be consequences of p; or [Kp & K(p -> q)] -> Kq.

He then proceeds to offer a demonstration of the claim that RA' and Closure are inconsistent: RA' embodies an epistemological externalism, while Closure is characteristically epistemologically internalist. To attribute to the content skeptic an argument that requires accepting both

externalist and internalist approaches to knowledge is to attribute to the content skeptic an incoherent position, which is uncharitable (cf. Brueckner, 1994, pp. 330-334).

However, he maintains, there is a reconstruction of the content skeptic's argument on which it does not appeal to RA', and is not incoherent. The proposal is to replace RA' with the following principle:

> U Suppose that I am considering an hypothesis H and a competing incompatible hypothesis SK. If my evidence and reasons (and whatever other considerations are available) do not favor H over SK, then I do not have justification for rejecting SK; hence I do not know that not-SK. (Cf. Brueckner, 1994, p. 333.)

Armed with U, the content skeptic can offer this argument:

(1) Externalism is true.
(2) Given externalism, S's evidence for believing that she is thinking that water is wet does not favor the belief that she is thinking that water is wet over the belief that she is thinking that twater is wet.
(3) Therefore, S has no justification for believing that she is not thinking that twater is wet.
(4) Therefore, S does not know that she is not thinking that twater is wet.
(5) Therefore, S does not know that she is thinking that water is wet.

The move from (2) to (4) is characteristic of internalism in epistemology. Closure is required to get from (4) to (5).[12] But, unlike the RA' argument, there is nothing contradictory in accepting these two moves together. Thus, Boghossian's argument need not be reconstructed in the way that F&O suppose; there is, according to Brueckner, a coherent argument from externalism to skepticism about knowledge of content, against which the standard strategy says nothing.[13]

Brueckner appears to believe, contrary to F&O, that this sort of skepticism is just fine (i.e., he seems not to share with practitioners of the standard strategy the worry that if externalism implies the absence of KC, it must be false). I agree with standard strategists that externalism is threatened to the extent that it implies the absence of KC; and while I agree with Brueckner that externalism, ultimately, does imply the absence of KC, I do not believe that Brueckner has said anything that is even remotely effective in undermining the standard strategy for establishing the compatibility of externalism and KC.

Hence, we cannot rely on Brueckner for a defense of Boghossian's argument.

One problem is that Brueckner is simply mistaken in thinking that F&O have canvassed his original (1990) content skeptical argument as a deployment of reasoning along the lines of RA'. F&O's argument based on RA' does not involve (c") and (d'). Thus, to the extent that Closure is required only to get from (d') to (d), it is not a part of F&O's RA' argument. It may be, of course, that Closure is somehow presupposed in the reasoning that leads to F&O's diagnosis of the RA' argument (i.e., that (c') is false), but it is at least not obvious that it does. Thus, Brueckner's reading of F&O's use of RA' is mistaken.[14]

As a matter of fact, F&O have merely applied their *diagnosis* of the difficulty with the RA' argument to Brueckner's content skeptical argument. They claim that what they (following Burge) say against RA' (i.e., the standard strategy) applies with equal force to Brueckner's argument: The fact that second-order thoughts inherit their content, including their environmental determination, from the first-order thoughts they involve, shows that (c') is false; in cases involving knowledge of content, S would *not* continue to hold the belief even if it were (owing to different environmental determiners) false (this is the first component of the standard strategy). The fact that S could not tell the difference between the thoughts she would think in the two different environments is irrelevant, since F&O reject CKC; knowledge of content cannot fail, even if comparative knowledge of content might (this is the second component of the standard strategy).

F&O believe that the standard strategy also shows that (c") is false, and that the antecedent of Closure need not be satisfied in the context of knowledge of content, thus blocking the move from (d') to (d), and from (4) to (5). There is no opportunity for the second-order thought to misidentify the first-order thought, so there is no opportunity for the second-order thought to be mistaken in the way (c") maintains (see F&O, 1994, pp. 122-3); S still knows the contents of her thoughts (according to the first component of the standard strategy), and this cannot be challenged by appeal to the fact that S could not tell the difference between her thoughts in the two environments (according to the second component of the standard strategy).

So, Brueckner's version of the argument does not establish that S does not know that she is not thinking that twater is wet. But even independently of this, F&O maintain that the rejection of CKC also exposes a weakness in the reasoning from (d') to (d) and from (4) to (5), supposedly sanctioned by Closure, without requiring one to reject Closure. The inference involves a *reductio* from the assumption that S knows that she is thinking that water is wet (K(w)), and the (now known to be

mistaken) assumption that it is proven that S does not know that she is not thinking that twater is wet (-K(-t)):

A.	-K(-t)	assumed/proven
B.	K(w)	assumed
C.	K(w -> -t)	CKC
D.	K(-t)	B,C (Closure)
E.	-K(-t) & K(-t)	A,D
F.	-K(w)	B-E

It is worth noting why CKC is taken as support for premise C:

F&O (1994, p. 120) claim that if one grants Closure, the rejection of CKC shows that the second conjunct in the antecedent of Closure (i.e., premise C in the above argument) might not be true:

> "One apparent problem with this argument is that it is not obvious that if I know that I am thinking that [water is wet], then I know that I am not thinking that [twater is wet]. Whether this is the case depends on how I am introduced to the concept twater. If twater is introduced to me as a liquid whose chemical structure is not H_2O, then it is possible that I know that I am thinking that [water is wet] without knowing whether this thought is the same as the thought that [twater is wet]. For, as we argued above, introspective knowledge of content [KC] does not involve or entail introspective [comparative] knowledge of...content. In such cases where I lack introspective [comparative] knowledge of...content, I can know that I am thinking that [water is wet] without knowing that I am not thinking that [twater is wet]. This is not to say that deductive closure fails to hold in such cases, but rather that the antecedent of the closure condition is not satisfied."

What they mean here, of course, is that it is the second conjunct of the antecedent of Closure that is not satisfied (they concede explicitly that the first is satisfied). In this passage, F&O actually argue against a claim that differs importantly from C, the second conjunct of the antecedent of Closure. F&O's remarks are directed in the first instance against the claim that

C'. K(w) -> K(-t)

They claim that C' might well be false even if Closure is true. The only circumstance under which C' could be false while Closure is true, however, is if C, the second conjunct of the antecedent of Closure, is false

(because within Closure, the second conjunct of the antecedent is the only thing standing between K(w) and K(-t)). F&O argue in the quoted passage that the rejection of CKC permits the rejection of C'. And if C' is false and Closure is true, it must be that C, the second conjunct of the antecedent of Closure, is not true. With the rejection of CKC, we can see how C might be false. S may fail to know that if she is thinking that water is wet, then she is not thinking that twater is wet, because she may not be aware of any differences between water thoughts and twater thoughts. For, as F&O argue, if she cannot tell the difference between water thoughts and twater thoughts, and she is introduced to twater in the right way (i.e., not as distinct from water but as distinct from H_20, assuming she does not know that water is H_2O), then she may conclude that there is not any difference between water thoughts and twater thoughts. On the other hand, if CKC is true, then S would know that her water thoughts and twater thoughts are different (provided she knows her water thoughts), and hence that if she is thinking a water thought, she is not thinking a twater thought.

So, if the standard strategy is correct, specifically the second component rejecting CKC, then the inference from d to d' and (4) to (5) is not sound; step C may under certain circumstances be false. This is in addition to F&O's contention that the standard strategy shows that premise A is also unwarranted. Thus, their diagnosis of the argument they considered based on RA' applies with equal force to Brueckner's reconstructed argument based on RA', as well as the argument based on U that Brueckner himself supplies. Brueckner's supposition that content skeptical arguments based on U will fare better than arguments based on RA' is mistaken; it is mistaken primarily because Brueckner offers nothing that even tends to impugn the standard strategy.

It should not surprise us that Brueckner's objection to F&O is not successful. He based his objection to F&O's attribution of RA' to the content skeptic on the grounds that RA' is epistemologically externalist, while Closure is epistemologically internalist. He then argued that the two are incompatible. But this argument fails to take account of the fact that the epistemological externalism of RA' is contained in its ability to rule out of court irrelevant alternatives. F&O have previously agreed, however, that whatever the notion of relevant alternatives come to, it is not likely to rule out twater thoughts in the context of questions of knowledge of the contents of water thoughts.[15] Thus, they essentially abandon appeals to epistemological externalism in this context. F&O (1994, p. 120ff) make this explicit when they discuss a version of the content skeptic's argument that parallels the epistemologically internalist interpretation of the Cartesian skeptic's argument against knowledge of the external world. In essence, F&O simply anticipated an

argument along the lines of Brueckner's principle U, and refuted it by deploying the standard strategy.

The claims made in (c) and (c'), accordingly, are fully consistent with epistemological internalism. Indeed, the inference from (c) to (d) in the argument based on RA is essentially equivalent to that found in the argument based on U; if U is consistent with epistemological internalism, then so is RA. Thus, if U is consistent with Closure, then so is RA. And since F&O move from RA to RA' only because they take the move from (c) to (d) to be underwritten by the move from (c') to (d), RA' will be consistent with epistemological internalism as well, and hence also consistent with Closure.

Brueckner's claim that RA' is inconsistent with Closure depends on reading RA' in a way that F&O do not. Consider Brueckner's assertion that

> "...on the (fairly standard) understanding of the notion of *relevant alternative* which Falvey and Owens accept, it is apparently possible that S should know that ø (and hence *not* satisfy the antecedent of the appropriate instance of RA'), know that ø entails ß, and yet not know that ß (in virtue *satisfying* the antecedent of the appropriate instance of RA')." (Brueckner, 1994, p. 330.)

This, obviously, is purported to be a description of the failure of Closure; Brueckner then goes on to produce an example that fits this description. But since F&O are prepared to grant that twin thoughts are relevant alternatives, Brueckner's argument is not responsive to their deployment of the standard strategy. F&O's argument against content skepticism should not be taken to depend on epistemological externalism, even though it clearly does go through on an externalist epistemology. If F&O are epistemological externalists, then they would indeed reject Closure; however, they are not obliged to reject Closure since they claim merely that in cases of knowledge of content, where CKC fails owing to externalism about content, the second conjunct of the antecedent of Closure will not be satisfied.

So, insofar as RA and RA' can be understood consistently with epistemological internalism, U is not an advance upon it. If the standard strategy undermines the moves in RA and RA', it will also undermine the deployment of Closure in the move from (4) to (5). To rebut the standard strategy, then, one would have to show either that second-order thoughts do not inherit their content from the first-order thoughts, or that neither this fact nor any other warrants the rejection of CKC. Brueckner has said nothing that even touches on these points. We may grant that, as far as considerations attaching to RA, RA', and U

are concerned, the standard strategy does not succumb to Brueckner's content skepticism. Thus, Boghossian's argument remains in need of some response to the standard strategy.

2.5 A Successful Internalist Rejoinder: Linking KC and CKC

In section 2.3, I represented the standard strategy as offering this response to Boghossian's appeal to relevant alternatives in the attempt to show that externalism is incompatible with introspective knowledge of the contents of one's thoughts: When forming a second-order thought that one is thinking that water is wet, the externalist faces no special problem in accounting for S's knowledge that it is the concept of water, as opposed to the concept of twater, that is deployed in the first-order thought. The second-order thought simply contains the first-order thought as a part; if S's history of exposure to water guarantees that it is a concept of water (not twater) that is deployed in the first-order thought, then that same fact guarantees that it is that same concept that is featured in the second-order thought. The *knowledge* that one is thinking that water is wet derives from the fact that, in cases of basic self-knowledge, error is impossible. Burge explains why basic self-knowledge is immune to error by appealing to the reflexive nature of basic self-knowledge. No counterfeit thoughts are possible since the second-order thought is not numerically distinct from the first-order thought; the former simply involves the latter as a component. Thus, knowing that one is thinking that water is wet is not a matter of comparing the water thought to thoughts with different concepts; KC does not entail CKC. KC is just a matter of thinking the first-order thought, and thinking a second-order thought that involves that same content, as Burge (1988b, p. 656) puts it, "in the same mental act".

The standard strategy at least seems to be invulnerable in application to so-called basic self-knowledge; it is certainly not vulnerable to Brueckner's criticisms. But weaknesses emerge in its application to non-basic cases. In discussing these weaknesses, it will emerge that the sort of self-knowledge that the standard-strategy preserves is so severely circumscribed as to be trivial. Any substantial self-knowledge will turn out to be impugned by externalism, to its detriment..[16]

The standard strategy rules out the possibility of counterfeit twin thought contents, and hence explains the authority of basic self-knowledge, by maintaining that the first-order thought is thought in the *very same* thought as the second-order thought. But when the second-order thought is about a *previous* first-order thought, that first-order thought is *not* a part of the second-order thought. There is then the possibility that one might think that one was thinking, say, a

twater thought when in fact one was thinking a water thought; the slow switching case is precisely the sort of situation where this might happen. Suppose S, unaware that a switch has taken place, wonders at t_2 whether her current thought (suppose it is a twater thought) is the same as the thought she had at t_1 (which, suppose, was a water thought). According to the standard strategy (see, e.g., Burge, 1988b, p. 659), S knows at t_2 what she is thinking at t_2, and knew at t_1 what she was thinking at t_1, but does not (at least not without empirical investigation) know at t_2 whether she was thinking the same thought at t_1 (though she could presumably believe, mistakenly, that she was thinking the same thought). Let us assume, as Burge apparently does, that this is not because S does not know what it is for two thoughts to have the same content (though we are, of course, granting for the sake of argument that S is not able to *tell* of two thoughts whether they have the same content). Rather, it must be because at t_2 S no longer knows what she was thinking at t_1.

But how can the standard strategist explain why S no longer knows at t_2 what she thought at t_1? One possibility, which Burge (1988b, p. 659) suggests but does not develop, is that S does not, or cannot, remember what she thought at t_1. I grant that people frequently forget what they once thought, but this provides no general account of S's failure to know at t_2 what she was thinking at t_1. To see this, consider a thought S had around the time of t_1 that involved concepts that would not have been affected, even by externalist lights, by the slow switch of environments; suppose, e.g., that it is a thought about aluminum, and aluminum is the same in the old and new environments. S is certainly not debarred from remembering this thought. Why, then, should she be debarred from remembering the water thought she had at t_1? The ability to remember what one was thinking may not be infallible, but there is no reason (at least no reason independent of externalism) to suppose that its fallibility has anything at all to do with a change in environments. The externalist must concede that S can fail to know at t_2 what she was thinking at t_1, and that this is not because she has forgotten what she was thinking at t_1.

One avenue the externalist might explore to explain (or explain away) S's failure to know at t_2 what she was thinking at t_1 is simply to acknowledge that externalism sometimes entails such failures, and maintain that it is in the nature of thought contents that it is sometimes not possible to tell of two contents whether they are the same (because it is sometimes the case that one can fail to know at a later time what one was thinking at an earlier time). Of course, it remains true that, in spite of these occasional lapses, we often do know what we have thought in the recent and remote past. So we remain in need of an account of our authority with respect to these contents. Though the standard strategy

itself is silent here (because we are not considering a case of basic self-knowledge), an externalist tempted by the strategy might claim that when we are right, it is because the second-order thoughts inherit the environmental determination of the first-order thoughts, even when the first-order thought is not simply included in the second-order thought; we then have knowledge of content in those cases, and lack it in slow switching cases where the second-order thought content is determined by a different environment (see, e.g., Lepore and Loewer, 1986, esp. p. 612; Heil, 1988; Stalnaker, 1991, p. 144).

But this sort of explanation is inadequate. In the first place, as Brueckner (1990) argues (see also F&O, 1994, p. 119, fn. 12), this explains how S can think a second-order thought with the right first-order content, but it does not explain the authority of S's second-order belief; it does not explain why it is *knowledge* of content. There is no guarantee against error (since S *might* change relevant environments), so such second-order beliefs may be true but only accidentally so. Quite apart from this difficulty, however, there remains a deeper problem. This account of non-basic self-knowledge aspires to explain a phenomenon that we have not the slightest reason to believe exists. If the only reason to think that we have lapses in our ability to know at a later time what we were thinking at an earlier time is that externalism entails as much, then we have just as much reason to say that this is so much the worse for externalism. For it is just an *ad hoc* stipulation that such non-memory-related lapses can take place, one that is motivated only by the conviction that externalism is true; it is therefore not compelling to someone who does not share that externalist conviction. To be plausible, the externalist must justify the rejection of CKC as a necessary condition on KC independently of externalism.

As we saw in sec. 2.3, Burge attempts such a justification by comparing knowledge of content to perceptual knowledge: As in perceptual knowledge, it is not required that one be able to discriminate the object of one's knowledge from all possible counterfeits; one does not, in general, have to rule out the possibility of fake barns to know that one is seeing a barn in the Wisconsin countryside, and one does not have to rule out the possibility of twin thoughts to know what one is thinking (see Burge, 1988b, p. 654ff).

But it is plausible to suppose, as the relevant alternative approach does, that one *does* have to rule out the possibility of *relevant* alternatives. If barn facades are prevalent in the Wisconsin countryside, then in order to know that one is looking at a barn, one must take care to insure that it is a real barn and not a barn facade that one is looking at (though, perhaps, one need not prove that one is not looking at a hologram, if holograms are rare in that region). In the case of self-knowledge that we have been considering, S's water thought is a

relevant alternative to her twater thought. And, though Burge and other standard strategists do not challenge it, I argued in sec. 2.2 (see also Ludlow, 1995) that slow switching cases may in general be prevalent, so the twin thoughts to which they give rise may themselves be prevalent. Burge's comparisons with perceptual knowledge leave intact the claim that S's knowledge of the contents of her thoughts must be such that she can at least discriminate relevant alternative thought contents.[17] Denying this leaves the externalist with only an *ad hoc* contention that it is possible that we no longer know what we once knew about the contents of our thoughts even though we have not forgotten what we once knew.

Thus, it is only when we think a thought in what Burge describes as 'a reflexive way' that we can be said to know what we are thinking. When it comes to thinking about thoughts we had only moments ago, the guarantee provided by the first component of the standard strategy is lost. Thus, while the standard strategy may explain, consistently with externalism, how we can know what we are thinking at the exact moment that we are thinking it, it cannot explain how we can know what we were thinking at any other moment than the exact moment that we are thinking it. Of course, if we know at t1 what we are thinking at t1, then we are likely to know a few moments later what we were thinking at t1 since a change of environment can only yield a change in concept after some (as yet unspecified) period of time. But if we are unaware of the switches in question, then we never know whether we are nearing a point at which our concept is going to change, so at any given moment, it could be that we are about to lapse into having a thought with no particular content (if we can even call that a thought at all -- see Evans (1982), or a thought with some other content. Thus, there is no reason in general to suppose that if we have basic self-knowledge (of a thought as we are thinking it) we will also enjoy knowledge of the thought in question indefinitely thereafter. The standard strategy therefore buys us, consistently with externalism, only the most limited sense of (basic) self-knowledge. It is therefore consistent with externalism that we could lack knowledge of the contents of any thought we are not now thinking. Thus, it strikes me as though the standard strategy preserves only a sort of trivial self-knowledge. As far as the standard strategy is concerned, externalism has this consequence: Whenever someone asks us what we were thinking a moment ago, or what point we were trying to convey in an early remark, we will not be in a position to know. These questions are not about the thought we are having now, so the standard strategy is silent. The contents of these thoughts are things we might be mistaken about; if slow switching cases are indeed prevalent, then our inability to rule out the alternatives to which they give rise exposes the possibility of error.

Thus, externalism opens the door to myriad failures of self-knowledge for which there is no plausible account.

To clarify the points we have just examined, let us review what has been covered so far in this section: We have seen that there is a tension among the following three claims.

(1) S knew at t_1 what she was thinking.
(2) S knows at t_2 what she is thinking.
(3) S does not know at t_2 whether her current thought has the same content as the thought she was thinking at t_1.

The standard strategy provides the externalist with an account of (1) and (2), but at the expense of engendering a commitment to (3). The tension derives from the fact there is no adequate explanation for (3). It cannot, in general, be accounted for by supposing that S does not know what it is for two thoughts to have the same content; even if she is not a philosopher she might know what it would be for her to think the same thought on two different occasions. Therefore, (3) can be true only if

(4) S does not know at t_2 what she was thinking at t_1.

That is why S is unable to tell whether the two thoughts have the same content. But it is in general implausible to suppose that one is debarred from remembering what one was thinking at an earlier time; therefore, the explanation for (4) cannot be that S has at t_2 forgotten what she was thinking at t_1. The externalist is in need of an account that explains, without appealing to a faulty memory, how S could fail to know at t_2 what she was thinking at t_1.

Here, despite Burge's claims to the contrary,[18] the externalist can offer only two possibilities: The first is just an *ad hoc* stipulation that it is in the nature of content that such mistakes can happen (in which case we lack comparative knowledge of content). But this stipulation is highly implausible intuitively, and the motivation for it is not independent of externalism; so the stipulation should simply be denied by someone not already convinced of the truth of externalism. The second is an appeal to an externalist epistemology, such as the relevant alternatives approach. This option, however, entails that externalism is incompatible with wide ranges of self-knowledge. Twin thoughts often are relevant alternatives, so there are arbitrarily many cases where one does not know what one was just thinking moments ago, and especially days, weeks, or months ago. All that one can know, according to externalism, is the almost tautological claim that one is thinking what one is thinking. Any substantial self-knowledge is beyond the reach of the standard strategy.[19] For externalists appeal to

epistemological externalism to explain how one can fail to know now what one did not know before. With this appeal comes the conclusion that the earlier thoughts are relevant alternatives that must be ruled out in order to have knowledge. Without this appeal, the externalist is left with only the *ad hoc* and unexplained stipulation that such lapses in self-knowledge sometimes occur.

Thus, we may conclude that if S has KC, then she must also have CKC, at least with respect to relevant alternatives. Since externalism entails the systematic failure of CKC in certain kinds of case (i.e., slow switching cases), it also entails the failure of KC in those kinds of case as well. Intuitively, the point here is this: While the standard strategy rules out the possibility of error in basic cases (even slow switching cases), it does not rule out the inability to distinguish relevant alternatives. To the extent that we are working within the relevant alternatives approach to epistemology, then, the standard strategy cannot reconcile externalism with knowledge of content. It is, of course, implausible to suppose that we do not know the contents of ordinary thoughts such as the thought that water is wet. Thus, externalism appears to have implausible implications with respect to self-knowledge.

2.6 Defending the Link

In this section, I would like to defend the conclusion just reached against an opposing view that has recently received attention. F&O (1994) have contended that, while KC is undeniable, there are reasons to reject CKC as a necessary condition on KC independently of externalism, thus blocking the link between KC and CKC for which I have just argued (see also Owens, 1991). They ask us to consider these two sentences:

(i) Nobody doubts that whoever believes that Mary is a physician believes that Mary is a physician.
(ii) Nobody doubts that whoever believes that Mary is a physician believes that Mary is a doctor.

Do these sentences express the same thought, or different thoughts? It is difficult to say. There is disagreement on the matter; Benson Mates would maintain that (i) and (ii) express the same thought, while Alonzo Church would hold that they express different thoughts. F&O claim that in order to settle the dispute,

> "...one needs additional information about the world we live in, the nature of our linguistic practice, the semantic theories that

> best represent that practice, and so on. While much of this information is logico-philosophical in character, it is not plausible that it can be acquired independently of a serious empirical investigation into linguistic practice. So far as we can tell there is nothing in commonsense psychology to suggest that introspection alone provides all we need to ground judgments of sameness and difference in the contents of our propositional mental states." (F&O, 1994, p. 113.)

F&O advertise this as an argument against CKC (as a necessary condition on KC) that is independent of the truth of externalism. As a matter of fact, however, it is neither successful as an argument against the necessity of CKC, nor independent of externalism. There is a *prima facie* difficulty in seeing how the argument even applies to CKC. CKC is a thesis about our knowledge of the contents of our thoughts (a psychological issue), but F&O's argument is primarily concerned with the linguistic meanings of the sentences we use to express the contents of our thoughts (a linguistic issue). Mates and Church appear to have different views about whether they can use (i) and (ii) to express the same thought. But CKC is not about whether one can use two different sentences to express the same thought. CKC is about whether an individual knows authoritatively whether two of his thoughts have the same content. The fact that two individuals might disagree about whether two sentences express the same thought has very little to do with whether a single individual can tell whether two of his thoughts are of the same type.

On this latter issue, the relevant issue, it is hard to see that Mates and Church disagree at all. Consider the following variation on F&O's example. Of a given thought at time t_1, Mates believes that (i) expresses it. Church, too, has a thought at t_1 that he believes (i) expresses. Now, at time t_2, Mates has another thought that he believes (ii) expresses; Church, too, has a thought at t_2 that he thinks (ii) expresses. Mates believes that (i) and (ii) express the same thought type; it follows logically from this (which is not to say that Mates himself carries out this inference) that he thinks that his thoughts at t_1 and t_2 are the same thought type. Church believes that (i) and (ii) express different thought types; it follows logically from this that he thinks that his thoughts at t_1 and t_2 belong to different thought types. There is no disagreement between Mates and Church, however, unless Mates would say that Church's two thoughts are the same, and Church would say that Mates' two thoughts are different.

Are we entitled to conclude from what we have seen that Mates would say that Church's two thoughts are the same, and that Church would say that Mates' two thoughts are different? Not at all. Suppose

that Mates really is thinking thoughts of the same type at t_1 and t_2. Mates' views about language determine what sentences of English he believes express his thought contents; these views explain, in part, why he believes that (ii) expresses the thought he thinks at t_2. But suppose that Mates *antecedently* believes the thought at t_2 to be of the same type as the thought he had at t_1. Then the fact that he is thinking the same thoughts, and that he thinks he is thinking the same thoughts, together with the fact that he believes (i) and (ii) are synonymous, explains why he believes that (ii) expresses his thought at t_2.[20] Church, on the other hand, has different views on the semantics of sentences. If Church could know what thoughts Mates is thinking at t_1 and t_2, he would either claim that Mates is mistaken in thinking that (i) is an expression of the thought that Mates had at t_1, or that Mates is mistaken in thinking that (ii) is an expression of the thought that Mates had at t_2. He would *not* claim that Mates is mistaken in thinking that his thoughts at t_1 and t_2 are the same. It is simply not true, according to our supposition, and Church's theory of language would give him no authority on that matter.

But, F&O will counter, Church's theory of language would give him authority over whether or not Mates is thinking the same thoughts *if* the contents of Mates' thoughts are simply identical to the linguistic meaning of the sentences Mates uses to express his thoughts; and in fact F&O assure us that they are, since Mates "knows perfectly well the thought that each sentence expresses". In that case, we cannot say that Mates' thoughts are the same in spite of the fact that he would use sentences with different meanings to express them. Here, Church's theory of sameness of content would allow him to say whether or not Mates's two thoughts are the same, even without having to know the contents of either thought. (That is, suppose, contrary to fact, that Church does not really know what a physician is. In that case, he might not know what (i) and (ii) mean. Nevertheless, he would know that they do not mean the same thing, and therefore do not express the same thought, owing to his theory of the sameness of linguistic meaning.) And because Mates, we are supposing, has a defective theory of sameness of meaning, his claim that his thoughts at t_1 and t_2 have the same content is liable to be overturned by Church's superior theory.

Thus, given (a) that the contents of Mates thoughts are identical to the linguistic meanings of the sentences used to express them, and (b) that Mates knows (by KC) that the thought he had at t_1 is expressed correctly by (i) and that the thought he had at t_2 is expressed correctly by (ii), and (c) that Church is right that (i) and (ii) are not synonymous and hence express different thoughts, it follows that (d) Mates would be wrong in thinking that his thoughts at t_1 and t_2 are the same, and

Church would be in a position to correct him on that. So, even if KC is true, CKC may be false.

Let me first point out that even if I were willing to grant that this argument shows that Church could have knowledge of whether Mates' two thoughts are the same without knowing the contents of the two thoughts, it would not follow that KC does not imply CKC. For if this is what the argument is about, it is no more relevant than the argument F&O actually give. Church's knowledge of Mates thought contents, and their sameness or difference, is certainly not direct or introspective, so it does not involve KC or CKC. We want to know if Mates' knowledge of the contents of his thoughts at t_1 and t_2, which is direct and introspective, entails that he would know whether or not the two thoughts are the same.

That aside, the argument might be taken to show that, because Church could correct him, Mates could lack comparative knowledge of content even while he does have knowledge of the contents of his thoughts at t_1 and t_2. But the soundness of the argument, particularly premise (b), is highly suspect. What assurance do we have that Mates is correct in thinking that (i) and (ii) express the contents of his thoughts? Surely KC does not justify that claim; there are many ways in which Mates might err in expressing his thoughts even if he knows what they are. We are told, however, that Mates "knows perfectly well the thought that each sentence expresses". But it is clear that this is only meant to assert that he knows what each sentence means. And the sense in which he *knows* this is surely fairly weak. For Church, too, is said to 'know' what each sentence means, and yet he disagrees with Mates about whether they mean the same thing. Whatever F&O might want to claim about comparative knowledge of content, it is patently implausible for Mates and Church both to know the meaning of these sentences and disagree about whether they have the same meaning; at least one of them must be mistaken about the meaning of at least one of the sentences.[21] So Mates' so-called knowledge does not entail that he can put his thoughts into words with no possibility of error.[22] It is quite possible that Mates' (assumed-to-be) erroneous theory of sameness of content can corrupt his efforts to express the contents of his thoughts correctly. As we noted above, if at t_2 Mates has a thought that he thinks is identical to the thought he had at t_1, he may nonetheless express it by (ii) rather than (i) under the (mistaken) assumption that (i) and (ii) are synonymous. Thus, we are still without warrant for the claim made in premise (b) of the argument.

F&O could simply insist, however, that Mates has expressed his thoughts accurately by using (i) and (ii). Then, because he knows what he is thinking, and he has a faulty theory of sameness of meaning, he errs in judging his two thoughts to have the same content. But such

insistence is not an argument, so if that is all that F&O supply in support of (b), there is no reason to take the claim seriously.

About the only claim that I can imagine supporting (b) is the claim that one determines the contents of one's thoughts by first determining the meaning of the sentences one uses to express them. For that, together with KC, would justify premise (b). F&O are claiming, after all, that one has to have the right semantic theory to judge sameness and difference of content. Perhaps this is because one has to have the right semantic theory to know the contents of one's thoughts. But while this is the procedure that we are supposing Church would use to determine the contents of Mates' thoughts, it is surely not the procedure that Mates uses to determine the contents of his thoughts. The whole point of KC is that Mates has *direct* access to the contents of his thoughts, not access that is mediated by inference, least of all inference involving semantic theories. But the same is surely true in the case of comparative content. Church would again determine whether or not the contents of Mates' thoughts at t_1 and t_2 are the same by the same procedure that he would use to determine the content of Mates' thoughts; i.e., attending to the linguistic meaning of the sentences Mates uses to express the contents of his thoughts. But, just as Mates does not determine the contents of his thoughts at t_1 and t_2 by first determining the meaning of the sentences he uses to express them, he similarly does not determine whether or not the contents of those thoughts are the same by first attending to the meaning of the linguistic sentences he uses to express them.

F&O's whole argument is misguided because it exploits a patently erroneous model of the procedure by which an individual assesses whether or not two or more of his thoughts are the same. They can only undermine CKC by making an assumption that, if applied to KC as well, would also undermine it. But not only is such an assumption implausible, it is pretty clearly externalist. If applied to KC, it would entail that the identity of thought contents depends on the meaning of the sentences used to express them, which are in turn dependent upon linguistic practice. If applied to CKC, it would entail that the sameness or difference of thought contents depends on the sameness or difference of the meaning of the sentences used to express them, which in turn rests on facts about linguistic practice as well. But F&O's argument was advertised to be independent of externalist assumptions. In fact, their argument is neither independent of externalism, nor plausible.

Perhaps, however, F&O only mean to claim that we can lack comparative knowledge of content, even while we have knowledge of content, because it is not clear what sameness of content amounts to. Mates seems to assume that if two terms refer to the same sort of thing, as "doctor" and "physician" do in this context, they mean the same

thing, even when interpreted in psychological contexts; the idea is that if one believes something about doctors, one believes something about physicians, since doctors are physicians. Church, on the other hand, seems to assume that, at least in psychological contexts, sameness of reference is not sufficient for sameness of meaning; the idea is that one might believe something about doctors, but not believe something about physicians, so long as one is not aware that doctors are physicians. Until we settle just what sameness of content consists in, it will be possible to know the contents of two thoughts directly, and yet not know whether they are the same or different. But in order to settle such a dispute, we will first have to determine the correct semantic theory, and that theory will have to be deployed in the determination of sameness or difference of contents. CKC, therefore, will be anything but direct, even while KC remains patently direct and authoritative.

There are two points in this argument. The first is that we cannot settle questions of sameness and difference of contents until we determine what is to count as sameness and difference. This is debatable, however. The fact of the matter seems to be that we have two senses of sameness and difference, neither of which is somehow more 'correct' than the other. Thus, it is not that we have to wait around for someone to tell us which semantic theory is correct in order to determine if we have CKC. Rather, we simply have to distinguish two senses of CKC, one corresponding to Mates' understanding, and one corresponding to Church's understanding. There are, then, two questions that need answering, one about whether we have CKC in the intensional sense, and one about whether we have it in the extensional sense. The fact that there are two senses of CKC does not mean that either one is only indirect and theory-dependent. The theory-dependence attaches to the questions (which sense of CKC is at issue), not the answers (whether each one is direct and authoritative).

The second point has to do with the answers. Do we, first of all, have CKC in the way Church would understand it? I have been arguing that we do. We determine the contents of our thoughts directly; that is what the authority inherent in KC consists in. And if Burge is to be believed, this knowledge is sufficient to identify our thoughts as the types of thoughts they are:

> "In basic self-knowledge [KC], a person does individuate his thoughts in the sense that he knows the thought tokens as the thought tokens, and *types*, that they are." (Burge, 1988b, p. 653, italics mine.)

If we determine what types of thoughts we are thinking independently of the linguistic meaning of certain sentences we would use to express the

contents of our thoughts, then we surely determine the sameness or difference of our thoughts independently of the linguistic meaning of the sentences we would use to express those thoughts, too. We simply identify the types to which the thoughts belong, and notice whether they are of the same type or different types; there is nothing more to comparative knowledge of content in Church's sense than that. Thus, if we have authoritative knowledge of content in this sense, then we would have grounds for claiming authoritative comparative knowledge of content in this sense as well.

But what about CKC in the sense that Mates would understand it? Can we tell whether the concepts used in two of our thoughts are extensionally equivalent? Determining extensional equivalence is often a tricky matter, and in the case of empirical concepts such knowledge is patently empirical. The concept of a creature with a heart, it turns out, is extensionally equivalent to the concept of a creature with a kidney. Our knowledge of the extensional sameness of these two concepts often is, therefore, patently indirect. But does this contrast with our knowledge of the extensions of concepts when such concepts are treated individually, rather than in comparison to one another? I think not. We may know quite directly what we are thinking when we think of a creature with a heart, but that is just to say that we have knowledge of, as it were, the intension of the concept. We may also know directly that *if* the concept has an extension, it would include all and only creatures with a heart. But we do not know by introspection whether the concept does have an extension; nor, of course, do we know by introspection whether the extension of one such concept is the same as another. (We might know by introspection whether the extensions of two concepts would be the same if they had extensions at all only if the concepts are not empirical concepts.) Thus, those cases where we lack introspective knowledge of comparative extensional content are just those cases where we would lack introspective knowledge of extensional content. The failure of CKC in this sense is matched by the failure of KC. But the question of relevance is whether CKC holds in cases where KC holds as well, not whether it holds in cases where KC is absent.

These points about extensions, however, are most likely moot. The issue in this debate is surely not whether we have knowledge of the extensions of our concepts. It is not knowledge of what our concepts apply to that we have directly, it is just knowledge of what concepts we have; and this must be understood independently of the extensions. Even semantic externalism purports to characterize content intensions. And, moreover, when understood intensionally, it would seem that if we have knowledge of the contents of our thoughts, then, as far as F&O have shown, we can also know whether they are the same or different. F&O have said nothing to remove the threat that externalism entails that

we do not have direct knowledge of content because they have said nothing against the claim that externalism entails that we lack comparative knowledge of content.

Thus, externalism seems to be inconsistent with introspective knowledge of content; if externalism is true, one would indeed have to know certain facts about one's environment in order to have knowledge of content, and in any case could not know those contents in any direct, privileged, or authoritative way. Apart from the attempts we have considered, I can think of no other way one might try to deny this inevitable conclusion. Boghossian is, after all, right to think that this follows even on an externalist epistemology in the relevant alternatives tradition. If F&O are right that this conclusion can form the basis for a *reductio* of (psychosemantic) externalism, then it would seem that we have set the stage for its eventual refutation.

3. *Internalism and Knowledge of Content*

Boghossian dampens any enthusiasm the foregoing conclusion might engender in internalists by arguing that internalism, too, is inconsistent with the facts about self-knowledge. Heil, too, advances an argument of this sort. In this section, I will present these cases, and respond to them.

3.1 *A Conflict Between Internalism and Knowledge of Content?*

The point that Boghossian wants to exploit is that

> "...according to he currently prevailing orthodoxy, even the internal (or narrow) determinants of a mental event's content are relational properties of that *event* (although they are, of course, intrinsic properties of the *thinker* in whom the events occur)." (Boghossian, 1989, p. 14.)

The relational properties in question are causal properties; according to what is sometimes called *functional role semantics*, the content of any particular state or event is given by its causal or functional role in a system of states or events. Water-thoughts have distinctive (actual and counterfactual) causal interactions with other thoughts. The problem this poses for knowledge of content is this:

> "To know that I just had a *water* thought, as opposed to a *gin* thought...I would have to know...that my thought has the causal role constitutive of a *water* thought, as opposed to one constitutive of a *gin* thought. But it doesn't seem possible to know a thought's causal role directly. ...[Y]ou cannot tell by mere introspection of an object that it has a given *relational* or

> *extrinsic* property. ...[Y]ou cannot know that an object has a given relational property merely by knowing about its intrinsic properties. ...[M]ere inspection of an object gives you at most knowledge of its *intrinsic* properties." (Boghossian, 1989, p. 15-6.)

Given this, it would indeed seem that internalism faces a problem similar to the one faced by externalism. Though one need not know facts about one's historical relation to the environment in order to have knowledge of content (as on externalism), it would nonetheless appear that one must make *some* inferences concerning the relational properties of one's thoughts in order to have knowledge of contents of those thoughts. Since knowledge of content has been assumed to be direct, even some inference is too much inference; internalism, too, conflicts with the facts about self-knowledge.

3.2 Internalist Responses

There are at least a couple of ways an internalist might want to respond to this argument:

One way, though not an entirely happy way, would be to deny that the internal determinants of content are relational. This, of course, raises a nest of difficult questions to which, Boghossian points out, there are currently no adequate answers. The absence of a plausible alternative to functional role semantics, however, may simply be a temporary stumbling block. It may even be irrelevant if contents cannot be reduced to some purely physical set of properties. It is only if one *assumes* that content must be accounted for in purely physical terms, that it is some complex function of physical properties, that one would be tempted by functional role semantics. If contents are instead non-reducible, non-physical properties of brain states, then it is not surprising that the inadequacies of functional role semantics have not suggested an alternative reductive account. The truth about which properties, intrinsic or relational, determine contents is currently not known; no one, not even externalists, has even the slightest idea how physical material can have semantic, intentional, or conscious properties. So, for all anyone knows, there *are* intrinsic and internal determinants (or supervenience bases) of content; and so, for all anyone knows, there is *no conflict at all* between internalism and knowledge of content. We *do* know, however, that there is a very *big* conflict between externalism and self-knowledge.

Heil has argued, however, that there remain problems for the internalist:

> "If the contents of one's thoughts depended entirely on the state of one's brain, for instance, why should that fact alone render our access to those contents any less indirect or problematical? Were it assumed that, in order to grasp the contents of our thoughts, we must ascertain the conditions that fix those contents, we should be out of luck if those conditions included complex neurological goings-on. Even if one insisted only on the obtaining of those conditions, it is by no means obvious why conditions that depend solely on instances of characteristics or events inside one's head should be taken to have an epistemological priority over those that depend partly on items or events occurring elsewhere." (Heil, 1992, p. 174.)

Heil here raises two points: First, the neural spaghetti of our brains is no more accessible to us without inference than the microstructure of water. If knowing the conditions that enable or support thought content is required to have knowledge of content, then internalism would seem to be in no better position than externalism. Second, knowledge of such enabling conditions is not required to have knowledge of content, so externalism would seem to be in no worse position than internalism. In both cases we merely need to think the first-order thought "self-consciously" (for Heil) or "self-ascriptively (for Burge).

But the internalist can borrow here from the externalists' appeals to epistemological externalism. Knowing that one is thinking that water is wet is not a matter of knowing, as Burge (1988b) refers to it, "the enabling conditions" for that content. Though we have seen that self-knowledge requires at least the ability to discriminate between relevant alternative thoughts, no reason has been given for thinking that we must also know what it is that fixes the contents of our thoughts. Thus, even if the internal determinants of content are relational, it may not be necessary to know those relations in order to have introspective knowledge of content. All that one must know is the content itself, and this is done by ruling out other (relevant alternative) contents one might think, not by knowing the supervenience base.

Externalism runs into difficulty not because self-knowledge requires knowing the determinants of thoughts, but because self-knowledge at least requires discriminating between relevant alternative thoughts; externalism entails that this is often not possible. Internalism, on the other hand, even if it must eventually concede that contents are determined by internal physical relations between concepts and/or thought contents, does not entail that we are unable to discriminate between relevant alternative thoughts. Differences in internal physical relations between the brain states that realize thoughts might yield different contents on internalism (just as different

external relations yield different contents on externalism). But those differences in content can perfectly well be introspectively detectable precisely because they result from internal differences that may differentially affect the brain mechanisms responsible for introspective capacities; external differences, on the other hand, do not affect such introspective mechanisms at all. We do not have to *know* what these determinants are (e.g., we do not have to know that it is this or that brain state that realizes a given content -- again, all that must be knowable is the content); we simply have to be able to *detect differences* in such realization conditions.[23] That is how the internalist can avoid the problematic cases involving differences in content that are not introspectively discernible.

The internalist can then question Boghossian's assertion that we do not have direct access to the causal profiles of mental events. Boghossian repeatedly claims that our introspective access is limited to discrete mental events; and we are encouraged (e.g., p. 15) to imagine that these events have no appreciable temporal extent. As a matter of fact, however, it may be that when we introspect on an occurrent thought, we access the event that constitutes that thought, as well as other events to which it is constitutively related. We can do this by introspection because the relations are internal. Such access may take time. Boghossian's own notion of cognitively substantial introspection might even help us to understand what this sort of introspection might amount to:

As we have seen, Boghossian claims that it is plausible to suppose that we know the contents of our thoughts, not just in the standard strategy's self-verifying way, but also in a cognitively substantial way. For example, we can *explicate* our thoughts; we can say what we have in mind in thinking the concept of water. Burge (1988b, p. 661-2, 1989) claims that such explicative abilities are not necessary conditions on having contents. This may be right, but the question at issue is whether such explicative abilities are necessary conditions on having *introspective knowledge of* contents. Dogs have thoughts without being able to explicate them, but their explicational inability, plausibly, is matched by an inability to form (knowledgeable) second-order thoughts. Knowing that one is thinking that water is wet may involve knowing what other concepts one's concept of water is constitutively or associatively related to, and what other thoughts one's water thoughts might causally interact with.[24] There is no particular reason to think that this knowledge is inferential. It involves, to be sure, a temporally extended episode, but that, in spite of Boghossian's tentative claims to the contrary (p. 15), entails nothing regarding inference; "direct" does not mean instantaneous. Cognitively substantial knowledge of the content of a water thought may simply be a

matter of introspecting a thought we identify by an internal representation of the word "water", followed by an assortment of related thoughts that subsequently come to mind as we continue to train our attention inward.

This suggestion allows us to make sense of Boghossian's assertion that knowledge of content comes in degrees, which Boghossian offers as evidence for the claim that self-knowledge, even basic self-knowledge, is cognitively substantial. We can explicate our thoughts to greater or lesser degrees, of course, and this may simply be because we have greater or lesser access, or devote greater or lessor amounts of attention, to the relational properties of the concepts involved in the thoughts we introspect. It would certainly be rare that such self-knowledge is complete, and a virtual miracle if it were infallible (since the temporal extent and volume of available information may tax storage and computational resources); but, of course, it is Boghossian himself who insisted that knowledge of content is neither complete nor infallible.

The defense of this picture of introspection does not rest on analogies of the sort Boghossian criticizes (cf. the example involving the purported inspection of the relational properties of a dime, p. 16), nor is it an *ad hoc* attempt to save internalism from trouble. Support for the view rests instead on two sorts of consideration: First, a denial of the unsupported assumption that introspection can target only isolated mental events, and operates in a direct fashion only instantaneously. Second, the claim that a rich notion of introspective access like this is needed to account for facts about introspection that Boghossian himself agreed are in need of explanation. In this light, it would appear that internalism is not so much threatened by the facts of self-knowledge as supported by them. I therefore reject Boghossian's suggestion that internalism faces the same sort of threat as externalism.

8.2 Heil and The Mind's Eye

Guided by the assumption that externalism is no worse off than internalism with respect to the facts of self-knowledge, Heil searches for the source of the internalists error in thinking they are better situation with respect to such facts. He suspects that the problem stems from the idea that knowledge of content is like perceptual knowledge; that there is a "mind's eye" trained inward on our mental contents. To combat this idea (something both Burge and F&O also try to do), Heil introduces an analogy with drawing. Tokening a thought content is like drawing a picture. When one draws a picture, one does not need to make very careful observations of the resulting drawing in order to know what the drawing represents (*ceteris paribus*). Rather, one knows what one has drawn simply in virtue of the act of drawing it. Tokening a thought content is analogous. When one thinks that water is wet, one need not

make very careful observations of that content in order to determine what it represents. Rather, one knows what one is thinking simply in virtue of the act of thinking it. This lends credence to Heil's (and Burge's and F&O's) idea that knowledge of content is just the self-ascription of the thoughts one thinks; if we can token an externally determined thought content, then that thought content is simply a component in the thinking, and knowledge, that we are thinking a thought with that content, in the same way that a figure may be a component in a larger drawing. Knowledge of thought content, then, is not a matter of looking at "phosphorescent" mental items and knowing what category they belong to.

Heil's analogy is compelling and, in a certain way, illuminating. What it illuminates, however, is not what Heil takes it to illuminate. I am prepared to agree that knowledge of content is as bound up with the tokening of that content as the drawing analogy leads us to believe. To deny that in thinking our thoughts we know what our thoughts represent is to deny what is obvious. I have, however, previously argued against the sufficiency of the self-ascriptive model. I do not think the drawing analogy lends appreciable support to that model, but just to the idea that knowledge of content is not an accidental feature attending our tokenings of content. It is, rather as the drawing analogy suggests, an intimate feature of the way in which we are able to token thought contents, though not, perhaps, an intimate feature of the way in which cats and dogs token their thought contents (more on which in a moment).

A drawing represents what it does because the drawer determines what the drawing represents.[25] That is how Heil proposes to explain the privileged status of one's knowledge of one's own content. Presumably, his thinking is this: Because one does not merely observe thought contents, the fact that internally determined thought contents are determined by intrinsic features of the thought token is no advantage for them over externally determined contents, whose determinants are not intrinsic features; that is, we could not tell just by looking at an externally determined content what that content is, but then, we do not tell what contents are just by looking, anyway. Instead, he maintains, we tell what contents are by, as it were, constructing them, like drawings. But it would appear that in constructing a drawing, the representational content of the drawing is determined entirely by the drawer, not by any features external to the drawer, and certainly not by anything of which the drawer is unaware; for if it were, then how could the drawer know just by drawing what the drawing represents? If thought contents are like drawings, then, they should be determined entirely by the subjects of the contents, not by anything external to them, and certainly not by anything of which the subjects are unaware; for if

they were, then how could the subjects know just by thinking what the contents are? If drawing is analogous to thinking, then it would seem to illuminate the respect in which thought contents are determined internally, not any respect in which they are determined externally.

4. Concluding Remarks

It seems to me to be fairly clear that externalism conflicts in deep ways with facts about knowledge of content. But given the complexity of the territory at the intersection of semantics, epistemology, and philosophy of mind, even a clear view can fall out of focus. To restore that focus, I have considered Boghossian's argument that, even on an externalist epistemology of the sort found in the appeal to relevant alternatives, (psychosemantic) externalism conflicts in deep ways with facts about knowledge of content. There is a lot riding on such an argument since those with externalist proclivities, such as Burge and F&O, are prepared to grant that externalism would have to give way in the event that it conflicts with knowledge of content. We saw that Warfield's discussion of relevant alternatives is ineffective against Boghossian's argument. Despite Brueckner's claims, however, the standard strategy does seem to pose a problem for Boghossian. It is only once we appreciate the connection between knowledge of content and comparative knowledge of content that we can see the shortcomings in the standard strategy. Thus, there is a problem for externalists who recognize the authority and directness of introspective self-knowledge. No similar worries threaten the prospects of compatibility between self-knowledge and internalism. At the end of the day, I maintain, internalism fares well with respect to our intuitions about knowledge of content, while externalism does not.

CHAPTER 3

SKEPTICISM

1. Preliminary Remarks

Suppose we take for granted the self-knowledge discussed in the previous chapter. What, then, is the logical relation between psychosemantic externalism and skepticism about the external world? Much recent philosophical attention has been devoted to this question.[1] Skepticism is the view that we do not know most of the things that, pre-philosophically, we think we know about the external world. It is sometimes supported by thought experiments that appeal to the evidential similarity between the normal case and the case where one is dreaming, one is deceived by an evil demon, or one is a brain in a vat. I do not believe that skepticism is true. But I do believe that the position is coherent, that it might be (or might have been) true. I also believe that there are legitimate limitations on what we can be justified in believing about the external world; if this is a form of skepticism, then I am a skeptic of this sort. The same sort of justificatory gap does not plague the epistemic access we have to the contents of our thoughts. As we will see below, I am not alone in these judgments; many others, including externalists, want to embrace the coherence of the stronger form of skepticism, as well as the truth of the weaker form, and yet maintain the privileged status of self-knowledge. The question to which this chapter is addressed is whether externalists are able to do what they want to do.

The argument that has inspired much of the recent discussion of the logical relationship between these views is due to Putnam (1981, ch. 1):[2] If externalism is true, then if S were a brain in a vat, S's utterances of the sentence "I am a brain in a vat" would not express the proposition that S is a brain in a vat. S's use of the words "brain" and "vat" (and the concepts she associates with those words) would not refer to a real brain or vat, just as, in a Twin-Earth-style thought experiment, the Twin Earthling cannot refer to water with his use of the term "water". If we accept that generally we have knowledge of the contents of our utterances, then, if S knows that her utterances of "brain" and "vat" mean brain and vat, it will be possible for S to know that she is not a brain in a vat. For if S has authority with respect to the contents of her utterances, then she can know that her utterance of "I am not a brain in a

vat" refers to brains and vats. Since S would not refer to brains and vats if she were a brain in a vat, on externalism, she could know by mere introspection that she is not a brain in a vat. In fact, S can come to know many arbitrary propositions about the external world simply by attending to the contents of her thoughts. For if externalism is true, she can argue, then certain empirical conditions must be satisfied if she is even to have the thoughts about the external world that she does; since she knows what she is thinking, she knows these empirical conditions are satisfied. Thus, if S knows that she is thinking about brains and vats, the she can know that she is not a brain in a vat (a BIV); for if she was a BIV, then she would not be thinking about brains and vats, but instead about electrical impulses or some such thing. The standard BIV argument against skepticism, therefore, cannot be correct. This line of argument can then be used to conclude more generally that skepticism is false; for the argument can be run with virtually any utterance or thought with empirical content. If S knows that she is thinking about trees, for example, then she can know that there are trees.[3]

Many philosophers sympathetic with externalism take this purported anti-skeptical implication to be too robust. They claim that if externalism (together with the self-knowledge assumption) implies that skepticism is false, and allows us to know empirical propositions simply by having introspective access to the contents of our thoughts, this "would clearly amount to a *reductio ad absurdum* of externalism" (F&O, 1994, p. 108). I think externalists are right to be suspicious of the potential anti-skeptical implications of externalism. Not only does it strike me, and many externalists, that knowledge of the contents of one's thoughts should not be sufficient for knowledge of the world, there is also a more general worry: The threat posed to skepticism by externalism is not just that skepticism would be false; if externalism is true, then skepticism would be *incoherent*. The fundamental nature of thought (i.e., its externalism) would rule out even the possibility that skepticism is true; we could not *think* about the external world without also being able to *know* about the external world.[4] As Nelkin (forthcoming, ch. 9) reminds us, however, we should be suspicious of any view that threatens to reveal that a venerable philosophical position is incoherent. Skepticism may well be false, as indeed I think it is, but I am reluctant to think that its falsity is a matter of necessity, that it can be ruled out, as it were, by definition. I agree with F&O when they say that skepticism involves epistemological issues that should be settled by our concept of knowledge, not by a theory of mental content.

Externalists who recognize this threat have tended recently to adopt responses that conform to a single general strategy. According to this strategy, the only versions of Putnam's argument on which externalism entails that skepticism is false beg the question by assuming

that skepticism is false. I will call this the QB strategy (where "QB" stands for *question-begging*). F&O (1994) offer the most pure form of this strategy; their argument has been endorsed recently by Forbes (1995, p. 217, fn. 21), and Brueckner (1992b) advances the key claim of the strategy. Burge (1988b, p. 655, fn. 6) and Heil (1992, pp. 164-172) also seem receptive to this strategy. Martin Davies (1994) defends a principle on the limitation of the transfer of epistemic warrant that also partakes in the strategy, though the question that gets begged in Davies version is somewhat different.

In section 2, I will lay out F&O's and Davies' versions of the QB strategy, and argue that there is no effective Putnamian response to it. There is, then, no *easy* refutation of the QB strategy. I will then argue in section 3 that the failure of Putnamian anti-skeptical argument should offer no comfort to QB strategists. The QB strategy shows that one cannot infer any empirical claims, and hence that skepticism is false, on the basis of externalism and plausible claims about self-knowledge without begging the question, but this result is insufficient to block the threat of F&O's *reductio*. The anti-skeptical implications that threaten externalism, I will argue, survive the deployment of the QB strategy.

In rough outline, the argument to that conclusion is this: Externalism is threatened by a *reductio* to the extent that it entails that introspective knowledge of content is sufficient for knowledge of empirical claims about the external world, and to the extent that it rules out even the possibility that skepticism is true no matter what account of knowledge turns out to be correct. QB strategists argue that it is not possible to draw an inference from introspective knowledge of content to knowledge about the external world, and hence refute skepticism, without begging the question against the skeptic. But even if such an inference is question-begging, it remains the case that, on externalism, introspection is sufficient for knowledge of the external world, and hence that skepticism must be false. The logical relation between externalism and self-knowledge, on the one hand, and skepticism, on the other, is untouched by the fact that the argument is question-begging. It remains true, in other words, that externalism and self-knowledge entail that we can have knowledge of the external world, and hence that skepticism is false, even if the inference from the former to the latter is question-begging. Thus, I will argue, the QB strategy fails to ward off the anti-skeptical implications of externalism that give rise to the *reductio*.

In a final section, I will consider Burge's attempts to address the relation between externalism and skepticism. I will argue that they, too, are unsuccessful.

2. *The QB Strategy*

I will begin by providing the context in which the QB strategy arises. I will then present the strategy in various forms, and defend it against quick dismissal.

2.1 *A Brief History: Arguments O and O'*

Brueckner (1986) offered this reconstruction of Putnam's argument, which he now calls argument O:

(1) Either I am a BIV (speaking vat-English) or I am a non-BIV (speaking English).
(2) If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are true iff I have sense impressions as of being a BIV.
(3) If I am a BIV (speaking vat-English), then I do not have sense impressions as of being a BIV.
(4) If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are false. [(2), (3)]
(5) If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are true iff I am a BIV.
(6) If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are false. [(5)]
(7) My utterances of 'I am a BIV' are false. [(1), (4), (6)]
(8) My utterances of 'I am a BIV' are true iff I am a BIV.
(9) I am not a BIV [(7), (8)]

Brueckner formerly criticized argument O on the grounds that (8) could not be known to be true in the context of an argument against skepticism. He now believes (at least as far as his 1994 paper is concerned) that this criticism is unwarranted (see also Dell'Utri, 1990, Hill, 1990, and Tymoczko, 1989). F&O agree, but nevertheless do not accept the argument. They insist that (2) and (3) have to be rewritten, so that the language (vat-English) will take account of an appearance-reality distinction. As they stand, (2) and (3) do not reflect a difference between introspectable sense impressions and the real causes of those sense impressions, some electronic configuration Q in the computer that is simulating the BIV's experiences. Thus, O must give way to O', which is identical to O, except that (2) and (3) become:

(2') If I am a BIV, then my utterances of 'I am a BIV' are true iff my electronic configuration is of type Q.
(3') If I am a BIV, then it is not the case that my electronic configuration is of type Q.

F&O object to O' on the grounds that there is no reason to accept (3'). (3) was acceptable on the grounds that it can be known by introspection that the truth condition for a BIV's utterance of 'I am a BIV' is not satisfied; the truth condition is the existence of the sense impressions, which are stipulated in the original BIV thought experiment to be the same whether one is a BIV or not. But in (3') the truth conditions become the normal causes of those sense impressions, and hence the utterance of 'I am a BIV'; and while we can know by introspection that the sense impressions will be the same whether one is a BIV or not, we cannot know by introspection whether the normal causes of those sense impressions will be the same whether one is a BIV or not. Therefore, according to F&O, the flow of argument represented in O' stops at (3'), a premise for which there is no support (see F&O, 1994, pp. 130-31).[5]

Brueckner responds to F&O's analysis by pointing out that their reservations about (3') attach only to the consequent of the conditional expressed in (3'), and not the conditional itself. The conditional claim may well be justified even if the consequent is not. And, he argues, the conditional can indeed be justified by appeal to Putnam's original BIV thought experiment. In that thought experiment, neither I nor the BIV ever utters 'I am a BIV'; neither one of us is ever caused to do so. Nothing in my environment or the BIV's environment counts as the normal cause of that utterance; *a fortiori*, electronic configuration Q does not count as the normal cause of the BIV's utterance of 'I am a BIV'. Indeed, the BIV's electronic configurations are never Q, because if they were, the BIV would (at least under normal circumstances) utter 'I am a BIV', which, by hypothesis, the BIV does not do (see Brueckner, 1994, p. 337).

Whatever the merits of Brueckner's response, F&O consider a more direct way to get at the basic points at issue. In what follows, I will devote my attention to the direct argument, though the points I make will apply with equal force to all versions of the Putnamian argument insofar as they all depend upon externalism (this includes the version of the argument that Brueckner, 1994, calls argument S, which is similar to the version discussed in Tymoczko, 1989).

2.2 *Argument S' and the QB Strategy*

F&O consider an argument that Brueckner (1994) calls argument S':

(A) I am now entertaining the thought that I am not a brain in a vat (BIV).
(B) A BIV could not entertain the thought that it is not a BIV
(C) I am not a BIV. (From F&O, 1994, p. 132.)[6]

The argument *appears* to be a good one. My knowledge of (A) is supported by the claim that I have authoritative knowledge of the contents of my thoughts. (B) is, of course, supported by externalism with respect to mental content. If externalism is true, then, I would be able to know the truth of (C), and all that follows from it, by mere introspection of the contents of my thoughts.

F&O believe that the argument does not go through. Their criticism leans heavily on the fact that the skepticism implicitly rejected in the conclusion of the argument is global in character; it is the claim that all of our beliefs about the external world are false. Because of this, the argument against it cannot rest on empirical assumptions, those that would be true only if the conclusion is true; they are the very claims at issue. But F&O point out that the standard arguments for externalism, and hence the arguments that would support (B), appeal to specific claims about the nature of the world. These arguments depend on establishing that the twins in a Twin-Earth-style thought experiment have different concepts, and that this difference is owing to differences in their environments, not in their own internal states (individuated non-intentionally and individualistically). In Putnam's original example, Oscar's utterances of 'water' refer to H_2O since that is what he has had exposure to in forming his concept; Twin-Oscar's utterances of 'water' refer to a substance XYZ for the same reason. XYZ is not H_2O, but water is H_2O, so XYZ is not water. F&O point out that the fact that water is H_2O and not XYZ is crucial to the case for externalism. But this is an empirical fact, precisely the kind to which one cannot appeal in an argument against skepticism. When applied to the anti-skeptical argument above, the question-begging nature of the externalists' support for premise (B) is apparent:

> "...[I]n order to establish that the BIV's utterances of the words 'brain' and 'vat' differ in reference from our utterances of these words, we must appeal to the premise that our words 'brain' and 'vat' refer to definite physical objects in our environment. But these are empirical propositions, which in fact are true only if we are not brains in a vat. As such, the appeal to this premise is illegitimate in the present context. It begs precisely the question at issue." (F&O, 1994, p. 135).

So, F&O claim, there is no argument from externalism to the rejection of skepticism that does not depend on empirical claims (claims that would not be true if skepticism were true).

In responding to F&O's argument, Brueckner (1994) draws attention to F&O's attempt to establish the truth of (B).[7] They appeal

to Putnam's original Twin Earth thought experiment as a model for the kind of support they eventually supply for (B). Brueckner claims that this model is inappropriate. Whereas both twins in the Twin Earth story causally interact with liquid substances, the BIV and the non-BIV in the present story do not both causally interact with liquid substances (in the case of water), nor do they both interact with "body parts of some sort" (in the case of brains).[8] If we make the "innocuous" assumption that water, if it exists, is a liquid substance and not an electronic impulse, and that equally "innocuous" assumption that brains, if they exist, are body parts of some sort and not electronic impulses, we will see that F&O have no basis for probing the support for (B), and charging that it begs the question against skepticism. These "innocuous" assumptions by themselves are sufficient to support (B) without begging the question against the skeptic.

The obvious rejoinder to this response, which Brueckner anticipates (but apparently does not appreciate), is to say that the purportedly innocuous assumptions are after all *empirical* assumptions, and hence *do* beg the question against the skeptic in exactly the same way as the empirical assumptions that F&O consider explicitly.[9]

Brueckner's response to this rejoinder purports to show that even it has the consequence that I am not mistaken in my belief that water exists. Interestingly, he leaves it up to us to draw the relevant conclusion, the one that applies to (B), that I am not mistaken in my belief that I am not a BIV. The idea is that if we cannot even assume that water and brains are not electronic impulses, then perhaps they are; that, at any rate, is what he interprets the resistance to his "innocuous" assumptions to come to. In that case, then, even if I were a BIV, I would not be mistaken in thinking that water exists; it is just that it would be a certain kind of electrical impulse rather than a liquid. It also follows from Brueckner's line of argument, that I would not be mistaken in thinking that brains exist, and hence that I am not a BIV, it is just that brains would be electrical impulses rather than "bodily parts".

But if that is so, then not being a BIV, presumably, would also amount to being a certain kind of electrical impulse. Brueckner takes this possibility to land us back in argument O', which goes through on the assumption that (8), the disquotation premise, can be known to be true. Since we cannot know whether (8) is in English or vat-English, we do not know what is being asserted when it is asserted in (9) that I am not a BIV. Either I am not a BIV, where BIVs are brains (understood in the English sense), or I am not a BIV, where BIVs are electronic impulses (which is what vat-English brains would be). But, regardless of which interpretation applies, I can know that the statement in (9) is true. To the extent that establishing the truth of (9) defeats the skeptic,

argument O' will have defeated the skeptic (see Brueckner, 1994, p. 342).

But even if S' collapses back into O', that hardly insulates the Putnamian argument from the QB strategy. The QB strategy claims that in the defense of externalism, one must make empirical assumptions that beg the question against the skeptic. To the extent that O' depends on externalism, then it, too, begs the question against the skeptic. Therefore, Brueckner cannot defend S', and resist QB strategy, merely by exposing it to be of a kind with O'. The weak link in both arguments is the dependence on externalism; one can only defend the externalist premise in a Putnamian anti-skeptical argument by assuming an empirical claim that begs the question against skepticism.[10]

2.3 Limitations on the Transfer of Epistemic Warrant

I would like to present another variant of the QB strategy, from Davies (1994), which extends significantly the reach of the strategy.[11] The unique feature of Davies' version of the strategy is that it focuses the charge of question-begging on the self-knowledge premise of Putnamian arguments conforming to S', not the externalist premise.

To appreciate Davies argument, let us first consider, as he does, a point originally due to Wittgenstein. Davies asks us to consider this inference:

(1) My friend in New York has a tree in his garden
(2) If my friend in New York has a tree in his garden, then the Earth exists
(3) The Earth exists

The Wittgensteinian reaction to this sort of inference, the reaction that has inspired Davies, is to deny that (1) and (2) together provide epistemic warrant for (3), even though the entailment expressed in (2) is known to be true. The key claim is that warrant cannot be transferred to the entailed proposition when believing that proposition is a precondition for our warrant for the entailing proposition and its very status as a warrant for the entailed proposition. Thus, in terms of the example above, believing that the Earth exists is a necessary precondition on my coming to believe (based on ordinary evidence) that my friend has a tree in his garden (indeed, I cannot even believe that I have a friend or that he has a garden without believing that the Earth exists). Therefore, my belief that my friend has a tree in his garden cannot be a reason for me to believe that the Earth exists. For as a reason, it is question-begging; I have to believe (3) if I am to be swayed by the evidence to believe (1). To *argue* from (1) to (3) is, in a certain

way, redundant; if you believe (1), you *already* believe (3), so there is no need to argue for it.

Based on this sort of consideration, Davies offers the following principle of limitation on the transfer of epistemic warrant.

> Epistemic warrant cannot be transferred from A to B, even given the *a priori* known entailment from A to B, if the truth of B is a precondition of our warrant for A counting as a warrant.[12]

Let us apply this principle to a case involving a BIV. Standard arguments for skepticism with respect to the external world hold that if I were a BIV, I could (and would) still believe that, e.g., I am sitting at a desk. Davies' principle allows for this possibility. It then shows how I could know that I am sitting at a desk, and at the same time deny that I know by inference from this fact that I am not a BIV. If I seriously entertain (i.e., believe) the BIV hypothesis, then I could not count my sensory evidence for the belief that I am sitting at a desk *as* evidence. In order to count it as evidence, I would have to reject the hypothesis that I am a BIV. Thus, if I am swayed by the evidence to believe that I am sitting at a desk, I must already have rejected the hypothesis that I am a BIV. I cannot then use my belief that I am sitting at a desk as a warrant for rejecting the BIV hypothesis.

With apparently minor modification, we can adapt Davies' limitation principle so that it applies to externalist interpretations of BIV cases. According to externalism, if I were a BIV, I would not still believe that I am sitting at a desk (since I would not be appropriately related to desks). So, externalism entails that if I do believe that I am sitting at a desk, I am not a BIV. Moreover, according to externalism, if I know that I believe that I am sitting at a desk, then I can infer that there are desks and that I am not a BIV (because if I were a BIV, I could not believe what I know myself to believe). But, for reasons similar to those that support the aforementioned limitation principle, it is a precondition of my believing that I am sitting at a desk that I already reject the belief that I am a BIV. For if I were to entertain seriously the BIV hypothesis, I would not know (nor even believe) that I believe that I am sitting at a desk. Thus, it is not that my knowledge that I believe that I am sitting at a desk, together with my externalism, allows me to infer (provides me with a warrant for believing) that I am not a BIV (or that desks exist, etc.). Rather, it is that in coming to believe that I believe that I am sitting at a desk, I implicitly reject the skeptical hypothesis that I am a BIV. If I know that I believe that I am sitting at a desk (as I do according to the self-knowledge assumption), I already reject the BIV hypothesis, and cannot use my knowledge of the content of my thought as a warrant for rejecting the BIV hypothesis.

Davies conjectures that the foregoing line of thinking instantiates a second principle of limitation on the transfer of epistemic warrant:

> Epistemic warrant cannot be transferred from A to B, even given an a priori known entailment from A to B, if the truth of B is a precondition of the knower even being able to believe the proposition A.[13]

It is implicit in Davies thinking, I suspect, that this second principle inherits some of the plausibility of the first. If the transfer of warrant from A to B is blocked if believing B is a precondition on being swayed by the evidence to believe A, then the transfer of warrant from A to B should also be blocked if believing B is a precondition on believing A. Though Davies stops short of offering this justification for the second principle, it seems plausible to me, and I shall grant that Davies has shown that, even if externalism is true, epistemic warrant cannot be passed from the knowledge (or belief) that one believes that one is sitting at a desk (or that water is wet, etc.), to the belief that there are desks (water, etc.), and hence to the belief that one is not a BIV. The QB strategy, I conclude, shows that Putnamian anti-skeptical arguments are doomed; externalism is not committed to there being a non-question-begging way to draw an inference from knowledge of the contents of our thoughts to knowledge of external world.

2.4 Revenge of the Putnamian Anti-Skeptic?

A Putnamian anti-skeptic might try to claim in response to the QB strategy that it shows only that the premises in the Putnamian anti-skeptical argument cannot be *rationally believed* without begging the question against the skepticism. In F&O's version of the strategy, the argument for externalism, and for (B) in argument S', represent our reasons for thinking that externalism is true. But externalism may yet be true independently of any reasons we could offer for it. And while we might not be able to endorse externalism rationally without committing ourselves to the rejection of skepticism, the truth of externalism, independently of whether we endorse it, does not beg the question against skepticism. It is only when we try to justify the assumption of externalism that we run into the question-begging problems pointed out by QB strategists. In the case of Davies' version of the strategy, externalism entails only that in forming beliefs about the contents of our thoughts we implicitly reject skepticism. If we consider the contents of our thoughts independently of our second-order beliefs about them, no question-begging occurs. Thus, externalism, together with the claim of

self-knowledge, entails that skepticism is false without begging the question against the skeptic.

This is, however, an argument that could be of no interest to the Putnamian anti-skeptic. For it claims only that there could be an argument from the conjunction of externalism and self-knowledge to the rejection of skepticism; but whether there is or not, the Putnamian could not *offer* such an argument, complete with substantiated premises, without begging the question against the skeptic (as the QB strategy showed). And if she offers the argument without substantiating the externalist premise or embracing the self-knowledge premise, there will be no need for the skeptic to accept the argument precisely because one premise is unsubstantiated and the other is not even believed by the proponent of the argument. So, under the proposal being offered by the Putnamian, the anti-skeptical argument remains an argument whose premises cannot be accepted without begging the question. As soon as the Putnamian tries to support the unfounded premise or embrace the other, she must beg the question in the manner revealed by the QB strategy. Thus, this resistance to the QB strategy is no resistance at all.

Note that this result leaves open the possibility that externalism, or an externalist premise, together with self-knowledge, or a self-knowledge premise, entails that skepticism is false (more on which below). What it does not leave open is whether Putnamian anti-skeptics can refute skepticism by an argument that features externalism and self-knowledge in its premises. They cannot.

To my mind, the QB strategy affords us a decisive refutation of anti-skeptical arguments (by, e.g., Dell'Utri, Tymoczko, Warfield, and others) that conform to the Putnamian strategy.[14] Even if one were to claim, naively in my view, that we have knowledge that externalism is true, that would do no good whatsoever in *establishing* the anti-skeptical conclusion; it would simply amount to the *assumption* of that conclusion.[15]

3. *Criticism of the QB Strategy*

QB strategists, apparently unlike Brueckner (1994) and other supporters of Putnam, believe that had externalism provided the basis for a sound argument against skepticism, it would have revealed itself to have overly powerful epistemological implications. In this section, I will demonstrate that, in spite of the fact that externalism cannot be deployed against skepticism without begging the question, its truth would preclude the truth of skepticism; externalism remains committed to the claim that knowledge of content is sufficient for knowledge of the world. I will conclude, then, that there are deep logical tensions between externalism and skepticism that expose externalism to the *reductio* with which QB strategists are concerned.

3.1 Externalism, Skepticism, and the QB Strategy

F&O show, I think correctly, that if the Putnamian anti-skeptical argument is to be construed along the lines of S', it begs the question against the skeptic. They then infer from this, albeit tacitly, that any logical tension between externalism and skepticism has been vanquished. But this is not true.

In its most general form, the question that S' poses is whether externalism would commit us to the claim that introspective knowledge of content is sufficient for knowledge of the external world, and hence entail that skepticism is false no matter what the correct account of knowledge turns out to be. The specific way in which S' raises this question is by purporting to show that, if externalism is true, knowledge of the external world could be inferred from knowledge of the contents of our thoughts. The QB strategy, I am prepared to grant, shows that this specific route to the more general logical tension is unsuccessful; externalism does not entail that introspection is sufficient for world knowledge, and hence rules out the possibility that skepticism is true, for *that* reason.

Nevertheless, externalism may entail that introspection is sufficient for world knowledge, and that skepticism could not possibly be true, anyway. Whether it does or not depends on the logical relations between externalism and self-knowledge, on the one hand, and skepticism, on the other. In terms of argument S', the issue is whether the premises entail the conclusion. If (A) and (B) entail (C), then S' would show that externalism, together with home truths about self-knowledge, does make introspective knowledge sufficient for world knowledge, and engender a commitment to the claim that skepticism could not possibly be true.

F&O's discussion of S', however, does not address this question. They instead examine ways in which one might argue for the truth of externalism, and hence premise (B). Putnam's Twin-Earth-style arguments in support of externalism, and hence (B), appeal to empirical claims that would not be true unless skepticism were false, and (C) were true. This, F&O claim, shows that the arguments like S' are question-begging. But arguments purporting to establish the truth of (B) have nothing whatever to do with the logical relations that hold between (A) and (B), on the one hand, and (C), on the other. More generally, the logical relationship between externalism and skepticism, in the context of authoritative self-knowledge, does not depend on first establishing the truth of externalism. The question is only this: If externalism *were* true, would that (together with one's knowledge of the contents of one's thoughts) entail that one knows claims about the external world, and hence that skepticism is false? This question concerns arguments *from*

externalism, not arguments *for* externalism. In addressing only ways in which one might support (B) and externalism, F&O fail to address this question.

There are ways in which F&O might take their discussion to bear on the logical relation between externalism and skepticism. Suppose they think that externalism can be true only if there is a Twin-Earth-style argument for it; and suppose it is true that (B) is true only if externalism is true. On these suppositions, externalism, and hence also (B), would entail the empirical commitments of the Twin-Earth argument. But these empirical claims would not be available in the context of the anti-skeptical argument. Therefore, externalism would entail commitments that beg the question against skepticism. In this way, F&O's discussion of (B) might be taken to bear on its logical relation to the conclusion of S'.

This line of reasoning suffers from multiple defects. In the first place, it is surely false that externalism is true only if there is a Twin-Earth-style argument for it. Arguments for information-theoretic forms of externalism involve no Twin Earth considerations. Moreover, the proposed bearing on the logical relation between externalism and skepticism would be cold comfort to QB strategists. Construed in this way, the argument avoids one implausibly strong implication of externalism by replacing it with an even stronger, and hence less plausible, implication. The threat of a *reductio* attaches to externalism not just insofar as it entails that we could infer (or have epistemic warrant for) a claim about the external world on the basis of our knowledge of the contents of our thoughts. Rather, the threat derives from the more general possibility that introspective knowledge is sufficient for world knowledge, and hence that the truth of externalism requires that skepticism is false no matter what the correct account of knowledge turns out to be. On the suggested revision of F&O's analysis, the threat encompassed in the inability to make sense of even the possibility that skepticism is true remains even without the assumption of introspective knowledge. The truth of externalism alone would guarantee that skepticism cannot be true. An argument from the truth of externalism to the rejection of skepticism begs the question against skepticism precisely because the truth of externalism entails claims that make skepticism false. If externalism *were* true, then those entailed claims would have to be true as well, and that would entail that skepticism would be false. But if externalism, when it is joined with home truths about self-knowledge, is threatened by implausibly ruling out the possibility that skepticism is true in complete independence of an account of knowledge, then it would seem to be even more severely threatened by implausibly ruling out the possibility that skepticism is true on its own, without assumptions about self-knowledge.

Thus, the suggested revision of F&O's argument is no way to protect externalism from the threat of the *reductio*.

F&O's deployment of the QB strategy contains a fundamental flaw, one of two that is inherent in the QB strategy (the other is discussed in the next section). Whereas the primary concern of defenders of Putnam's argument is the fate of skepticism, the primary concern of QB strategists is the fate of externalism. Putnamian anti-skeptics are prepared to assume that externalism is true; they then argue that this premise (together with the assumption about self-knowledge) contributes to a sound argument against skepticism. QB strategists, on the other hand, do not assume that externalism is true; but they do attempt to mitigate its anti-skeptical implications to protect it from the threat of the *reductio*. When discussing Putnamian anti-skeptical arguments, however, QB strategists tend to lose sight of their own primary concern (with the fate of externalism) and instead temporarily adopt the primary concern of the Putnamians (with the fate of skepticism). They address the question of whether there is a non-question-begging way to support externalism. Putnamian anti-skepticism may be vanquished by a negative answer to this question, but the threat of the *reductio* is not. The status of arguments for externalism is irrelevant to the status of arguments from externalism; any anti-skeptical implications externalism might have are independent of its evidential support. And it is these anti-skeptical implications that pose the threat to externalism.[16]

The difference in the primary concerns of Putnamian anti-skeptics and QB strategists can be brought out by an apparent difference in the conclusions of the arguments each is considering. Putnamian anti-skeptics are often concerned to rebut skeptical arguments that appeal to the evidential similarity between BIV and non-BIV situations. Refuting that argument does not entail that skepticism is false, but only that one sort of argument for skepticism is unsuccessful. That skeptical arguments that appeal to BIVs are unsuccessful, however, hardly entails any absurdity for externalism since the failure of one kind of skeptical argument does not render skepticism incoherent, nor does it allow that knowledge of content is sufficient for knowledge of the world.

QB strategists, on the other hand, are centrally concerned with arguments that seek to show that the conclusion to skeptical arguments, skepticism itself, cannot be true. QB strategists are concerned about the possibility that externalism entails that knowledge of content is sufficient for knowledge of empirical claims about the world. This, I think, is the right way to view the anti-skeptical arguments that externalism is taken to ground. When Putnamian anti-skeptics claim to refute skeptical arguments that appeal to BIVs, they do so by claiming that we can know that we are not BIVs. But the same reasons that are

supposed to show that we are not BIVs would also show that there is water, that sofas exist, or any number of standard empirical claims. Thus, if externalism succeeds at all against skeptical arguments, it is capable of grounding an argument against skepticism itself. That is, externalism would be inconsistent with skepticism. QB strategists are concerned about this, and that is why they want to undermine anti-skeptical arguments grounded in externalism.

The point I have been insisting on, however, is that the inconsistency between externalism and skepticism survives the sort of criticism that QB strategists have offered. The question-begging nature of the support for externalism does indeed undermine the kind of argument Putnamians have used to try to refute skeptical arguments that appeal to BIVs, but it does not remove the logical tension between externalism and skepticism that gave rise to the *reductio* to which QB strategists are responding. Thus, as far as QB strategists have shown, the *reductio* stands.

3.2 Externalism, Skepticism, and Self-Knowledge

The limitation principle, however, might be taken to avoid this problem. For an interesting feature of the principle is that it blocks the inference to the rejection of skepticism even if it is assumed that externalism could be known *a priori* (Davies assumes that it could be known *a priori*). Perhaps, then, the logical conflict between externalism and skepticism can be vanquished by the limitation principle, since the principle does not presume that externalism is supported by question-begging *a posteriori* assumptions.

This line of thinking illustrates the second fundamental flaw in the QB strategy. The QB strategy addresses only the limited question of whether externalism entails that one can come to know facts about the external world on the basis of knowledge of the contents of one's thoughts. In the previous section, I alluded to the fact that this is not the only way in which a conflict between externalism and skepticism can arise. Externalism can be committed to the claim that knowledge of content is *sufficient* for knowledge of the world without thereby providing a *basis* for knowledge of the world.[17] The QB strategy, particularly the variant of it that employs the limitation principle, conflates these two claims. I am granting that Davies argues successfully that, even given externalism, one cannot know facts about the external world on the basis of one's knowledge of content. But externalism may yet be committed to the claim that knowledge of content is sufficient for knowledge of the world. If I know that I am thinking that water is wet, then, according to externalism, I can (and perhaps do) know that water exists. It is just that the antecedent does not provide my grounds for the belief expressed in the consequent;

however, the fact that the antecedent does not provide (epistemic) grounds for the consequent does not destroy the logical relationship shown to exist between externalism and the negation of skepticism. Thus, to the extent that it is a *reductio* of externalism that it rules out even the possibility that skepticism is true no matter what the correct account of knowledge turns out to be, the limitation principle cannot save externalism.

How is it possible for the conjunction of externalism and self-knowledge to guarantee the falsity of skepticism without the rejection of skepticism being substantively inferred from the conjunction of externalism and self-knowledge? One possibility is that it is just in the nature of the contents of self-knowledge that they are "world-involving", in which case it is simply in the nature of introspection that it involves access to information about the world. Many externalists have taken up a view of this sort, of course, and I have argued against this view in the previous chapter. If, however, the externalists are right about the compatibility of externalism and the direct, *a priori* nature self-knowledge, then my criticism of the QB strategy leaves them with implausible implications regarding the relation between introspective knowledge and world knowledge, as well as more general implausible implications regarding skepticism.

Another way to explain how self-knowledge could be sufficient for world knowledge, however, is that, if externalism were true, our knowledge of content would somehow have to be the product of our knowledge of the environment; knowledge of content, then, would be inferential and empirical in nature.[18] Many philosophers, including externalists, find that this implication is every bit as undesirable as the anti-skeptical implication; knowledge of content is simply *not* inferential and empirical, and any view that cannot account for its direct and *a priori* character must be rejected. In this case, then, externalism avoids the anti-skeptical *reductio* by saddling itself with a self-knowledge *reductio.* That is, externalism avoids the implausibility that stems from an *a priori* knowledge of the world, and hence an *a priori* rejection of skepticism, by embracing the implausibility that stems from an *a posteriori* account of self-knowledge.[19]

Let me summarize the line of thinking I am pursuing: Suppose QB strategists are right that any anti-skeptical argument involving externalism as a key premise will beg the question against skepticism, either in its support for externalism (as F&O maintain) or in forming the second-order beliefs that constitute knowledge of content (as Davies maintains). And suppose that I am right that this leaves intact the logical conflict between the two views, and that this exposes externalism to an anti-skeptical *reductio.* The explanation for the

entailment from externalist and self-knowledge premises to an empirical, and hence anti-skeptical, conclusion cannot be that knowledge of content provides the sort of evidential or epistemic basis for knowledge of the world that would support a substantive inference (the transfer of warrant) from the former to the latter (and hence the refutation of skepticism); the QB strategy rules that out. The most plausible remaining candidate explanation for this incompatibility is that knowledge of content *presupposes* knowledge of the world. But in that case, it is difficult to see how self-knowledge could be *a priori*. Thus, if QB strategists are right about the inference from externalist and self-knowledge premises to an empirical and anti-skeptical conclusion, and if the externalist and self-knowledge premises entail the empirical, anti-skeptical conclusion nonetheless, then externalism is also in danger of entailing that self-knowledge is not direct and *a priori*. Thus, the results of this chapter appear to conspire with the results of the previous chapter against externalism.

From the point of view of opponents of externalism, this result would kill two birds with one stone. The first is that in spite of the efforts of QB strategists, externalism entails that introspective knowledge is sufficient for world knowledge, and hence renders skepticism incoherent. Therefore, externalism faces the threat of a *reductio* based on its anti-skeptical implications; it makes introspection sufficient for world knowledge, and hence cannot make sense of even the possibility that skepticism is true. The second is that given those anti-skeptical implications, and the fact that they cannot be explained by the transfer of epistemic warrant from introspective self-knowledge to empirical beliefs about the external world, they may only be explainable by claiming that knowledge of content presupposes knowledge of the world. Therefore, externalism faces the threat of a *reductio* based on its implication that self-knowledge is inferential and empirical; it cannot make sense of the special *a priori* nature of self-knowledge.

4. Beyond the QB Strategy

Burge (1988a) anticipated something like the *reductio* we have been considering in this chapter. Unlike the QB strategists, however, he *has* argued that the validity of the move from externalism to the rejection of skepticism is suspect. In particular, he endeavors to establish that externalism is consistent with skepticism, and hence cannot imply the falsity of skepticism.[20] He contended that the argument that seeks to burden externalism with the absurd implications trades on

> "...conflating questions of counterfactually evaluating one's thoughts with questions of what thoughts one would think if one were in the counterfactual situation." (Burge, 1986, p. 124.)

Once we see this point, we can see that certain other forms of skepticism, though inconsistent with externalism, are actually incoherent; it is only the making of the conflation that gives the impression they are coherent. Moreover, certain other forms of skepticism could indeed be true consistently with externalism.

I think that Burge is right that certain forms of skepticism are consistent with externalism, but his arguments against the coherence of certain other forms of skepticism do not convince. We may grant that if externalism is true, then our thoughts would not be as they are if we were under the lifelong influence of an evil demon; we would not be having water thoughts because there would be no water for us to interact with, and hence no interactions on which to base our formation of the concept of water. Hence, the skeptical hypothesis that we are the lifelong victims of an evil demon's deceptions, or lifelong brains in a vat, is impossible. Nevertheless, our actual thoughts, the thoughts we have if things are the way we take them to be pre-philosophically, radically misdescribe the world as it would be if we were under the lifelong deceptive influence of an evil demon. Thus, while our counterfactual (i.e., demon-caused) thoughts would be radically different and true, according to externalism, our actual thoughts are quite mistaken when evaluated for truth-value against the counterfactual world; they are, as we might say, *counterfactually* false. They are not *actually* false, of course, because their actual truth-value depends on evaluation against the actual world, not the counterfactual world. Indeed, if our only reason for thinking that our actual thoughts could be radically false is the possibility that we are the lifelong victims of systematic deception, then, according to externalism, there would be no possibility that skepticism is true; externalism, in other words, would preclude the truth of skepticism. This, of course, would open externalism up to F&O's *reductio*.

But the possibility of lifelong deception at the hands of an evil demon is not the only reason we can give for skepticism. Suppose that an evil demon were *suddenly* to eliminate or alter the external world, while retaining our sense impressions of it. We would, in this counterfactual situation, have all the concepts we currently have, and yet have radically false beliefs involving those concepts. Externalism is patently consistent with skepticism of this sort. Thus, there is no logical tension between the two, and the threat of the *reductio* is vanquished.

It would be prudent to ask, however, whether there is any independent reason for thinking that it is only this "temporary" skepticism that could be true. For if there is no relevant difference between the temporary and lifelong varieties of skepticism, then to the extent that externalism is inconsistent with the lifelong skepticism, it is open to the threat of F&O's *reductio* after all. It is my view that if there is an intuition supporting the coherence of the more limited skeptical hypotheses, that intuition will support, with the same force, the coherence of the more radical lifelong deception. There is surely nothing more bizarre about a demon having deceived us for a long time rather than for just a short time. Indeed, at what point would we pass from skepticism that is consistent with externalism to skepticism that is not? Were we suddenly to come under the sustained influence of the demon, how long would it be before we would develop new and true beliefs about our new demonesque environment? There seems to be something unprincipled in the distinction between these two kinds of skepticism, and yet the externalist is firmly committed to the coherence of one and the incoherence of the other. There is something odd in this picture, an oddity that is related to the counterintuitive externalist consequence that our concepts could shift in meaning without us being aware. Thus, I submit, there is a clear intuition that if the temporary skepticism is coherent, so too is the lifelong variety.

As far as I can see, the only reason one would deny such an intuition is that one is antecedently committed to externalism. But that is to beg the question at issue. In the context of a dispute over a *reductio*, like the one we are considering about externalism, one cannot deny the absurdity of the purported consequence on the grounds that it follows from (and is hence supported by?) the assumption whose reduction to absurdity is at issue. If Burge is to convince us that lifelong skepticism is somehow incoherent, and hence that its inconsistency with externalism is not an embarrassment to externalism, he will have to do so on grounds independent of externalism.

5. *Concluding Remarks*

Burge has accepted the responsibility of removing the logical tension between externalism and skepticism; but it is a responsibility that has not been discharged. As far as Burge has shown, then, externalism is open to the threat of a *reductio* stemming from its implications with respect to skepticism and knowledge of the external world. The QB strategy, no matter which form it takes, has been shown to be ineffective in removing the threat of this *reductio*. The epistemological implications of externalism must now be acknowledged to be too robust for its own good. At a minimum, we should regard externalism with suspicion based on its epistemological implications.

Part II

Anti-Individualism and the Cognitive Sciences

CHAPTER 4

COMPUTATIONAL VISION THEORY

Our discussion so far has considered whether the propositional attitudes of folk psychology are individuated by appeal to factors in one's social or physical contexts. And, though we are now in possession of reasons to be skeptical of such claims, it remains to be seen whether the scientific treatment of psychological states warrants a similar rejection of externalism.[1] In this chapter, I will begin to assess the plausibility of anti-individualism in the context of scientific psychology by considering the individuative implications of computational theories in psychology, in particular, computational theories of vision.

1. Preliminary Remarks

A great deal of philosophical work has addressed the question of whether Marr's computational theory of early vision is individualistic.[2] Burge (1986a) has argued that, according to Marr's theory, visual states are individuated non-individualistically (see also Davies, 1991, who agrees with Burge, and Egan, 1991, who has argued that while individuation of computational mechanisms is individualistic, the individuation of visual content is not). Segal (1989, 1991) has denied that Marr's theory has these non-individualistic implications. More recently, Shapiro (1993) has argued that the entire debate has been misguided.

Shapiro observes that philosophers usually endeavor to determine whether psychological states are individuated individualistically by considering whether molecularly identical but environmentally distinct twins, call them S and Twin S, would be attributed psychological states with different contents. The question pertaining to Marr's theory, then, is whether it prescribes the attribution of visual states with different contents to S and Twin S; if it does so prescribe, then it fails what Shapiro calls the "T-I test" (the type-identity test) for individualism, and therefore would be revealed to be non-individualistic. Shapiro argues, however, that

> "the T-I test is not an accurate measure of a theory's commitment to individualism, and so appeals to molecularly identical twins

> should not, by themselves, carry the day in discussions of individualism." (Shapiro, 1993, p. 490.)

Thus, he claims, neither Burge nor Segal address the issue of individualism in discussing Marr's theory. Burge and Segal both purportedly fail to understand what grounds ascriptions of content on Marr's theory, and as a result involve themselves in a dispute quite independent of the issue of individualism. Once this aspect of Marr's theory is properly understood, Shapiro claims, it becomes apparent why the T-I test is a poor measure of a theory's commitment to individualism. This conclusion, if it can be sustained, would be of paramount significance, since it would entail that virtually every discussion of individualism in the literature (i.e., those that involve the T-I test) would be misguided in ways similar to Burge's and Segal's discussions of Marr's theory.

I believe, however, that Shapiro is mistaken in a fairly deep way, attention to which will allow us to raise and clarify several important issues involved in discussions of individualism. Shapiro's main mistake, shared by Burge (but, significantly, *not* by Segal), is to misinterpret the essentially *epistemological* nature of Marr's approach to content assignments. As a consequence of this, I will argue, there is no reason to take Marr's assignments of content to bear on the question of individualism. Shapiro and Burge, on the other hand, construe Marr's assignments of content as offering a solution to the *metaphysical* problem of individuating visual states. In fact, Marr's theory says very little about whether contents are individuated by factors internal to the perceiving subject or are instead (or in addition) individuated by factors in the subject's environment.[3]

However, contrary to Shapiro's main claim, I will argue that Marr's theory *could* have a bearing on the question of individuation by means of application of the T-I test; nothing in Shapiro's discussion even tends to show that the T-I test is an inappropriate vehicle for the discovery of individuation conditions for mental states. Thus, Segal's use of the T-I test to motivate his so-called *liberal* interpretation of content assignments on Marr's theory, on which twins are assigned the same content by the theory, is not threatened by Shapiro's claims.

Segal's liberal interpretation of content assignments, however, may generate difficulties that an individualist would do well to avoid. I will offer in section 5 of this chapter an alternative response to Burge's argument that accepts the assignment of contents on the so-called *restrictive* interpretation, but blocks Burge's argument to the conclusion that Marr's theory favors an individualistic individuation of visual states. I will therefore be defending what Davies (1991, p. 463) calls "conservative individualism".[4]

Marr's theory is an interesting testing ground for questions of individuation in psychology. As Burge (1986a, pp. 25-6) says,

> "The theory of vision maintains a pivotal position in psychology. Since perceptual processes provide the input for many higher cognitive processes, it is reasonable to think that if the theory of vision treats intentional states non-individualistically, other central parts of cognitive psychology will do likewise. Information processed by more central capacities depends, to a large extent, on visual information."

Moreover, as Segal (1991, p 485) notes, the restriction to a fairly well developed example of a computational theory in psychology assures us that our philosophical ruminations will be "informed of and constrained by the facts". Thus, there is much to be gained from a proper analysis of the individuative implications of Marr's theory.

2. *What is Individualism?*

The argument of this chapter, perhaps more so than others, depends on a careful understanding of individualism (internalism). To avoid confusion and deter potential objections, I will begin by emphasizing and clarifying what I take to be at issue in this debate.

2.1 Individuation

Recall that individuation is about classification; it is about how tokens are classified as being of the kinds that they are (and hence how kinds are distinguished from each other). As Burge (1986, p. 4) says, correctly individuating mental kinds is about stating how the natures of mental kinds are fixed. Thus, to individuate a token state is to specify the facts that make it true of that token that it is a member of a given kind.

Marr's theory makes *some* claims about the sort of properties that individuate visual states. Since Marr's theory is a computational theory, it classifies visual states in part by their computational properties. Most commentators also maintain, rightly in my view, that Marr's theory classifies visual states by their intentional properties as well.[5] Since it is uncontroversial that formal properties are individualistic, the debate over individualism has focused on the intentional properties of visual states.[6]

If token visual states have the intentional properties that they have solely in virtue of factors internal to the individual, then they are correctly individuated individualistically. Individualism about visual states can then be expressed in this way: The facts that make it true that a visual state V has content c are facts about an individual

considered independently of her environment. Anti-individualism, then, can be expressed this way: The facts that make it true that a visual state V has content c include facts about an individual's environment. These characterizations are consistent with other formulations of the views in terms of supervenience. Indeed, even Burge (1986a, p. 4) allows that individualism "...simply makes a claim of supervenience...".[7] More specifically, it makes the claim that the intentional properties of psychological states supervene on local properties of brains and brain states. Anti-individualism is not committed to the rejection of supervenience; it is simply committed to the rejection of *local* supervenience.

Misunderstandings arises when the individuation of mental states by appeal to content is confused with the individuation of contents. Individualism in psychology is *not* about what makes a content-type the type that it is; it is not about specifying what properties a content must have if it is to be the content that it is (rather than some other content). A content is itself a property of a mental state; the property of representing (or misrepresenting) something as being a certain way. There is no question to answer about whether the property of representing x as F is determined to be what it is, rather than (say) the property of representing x as G, by factors internal to an individual who has a state with that content. What makes the property of representing x as F *that* property rather than another is simply that it represents x as F rather than some other way. This much has nothing to do with what is internal to an individual and what is not. There is, however, the *further* question of what determines why (or what makes it the case that), e.g., some state of an individual's brain has this property rather than another, why it represents x as F rather than G; *this* question does have very much to do with what is internal to the individual and what is not. To be individuated individualistically, the state must have the content x is F (i.e., the content that individuates the kind *x is F*) in virtue of properties internal to the individual.

2.2 *Identification*

We can contrast the individuation of kinds with the mere *identification* or description of kinds. It is possible to identify a kind by properties that are coextensive with those that are required for its individuation, but that are not themselves properties in virtue of which a kind is individuated. Gold, for example, might be identified or even described by its color and consistency, but it is individuated (or individuated correctly) by its atomic number. Grandfathers might normally be identified or described by the appearance of their skin and hair, but they are individuated by their relations to their offspring's offspring. The distinction between individuation and identification is

similar to one in evolutionary biology between selection *of* and selection *for* a particular trait (see Sober, 1984). A sieve might select particles *for* their size, but if all the appropriately sized particles in that domain are red, the selection by the sieve will also be a selection *of* red particles. The particles that pass through the sieve are of a kind (able to be passed through the sieve) in virtue of their size; they are therefore individuated as that kind by their size. But since, *per accidens*, all the particles of that kind are also red, tokens of that kind may be identified by their color; they are not, however, individuated by their color.

This notion of identification applies to the debate over individualism, particularly as it concerns Marr's theory, in the following way. An individual may token a state with the content x is F in virtue of factors internal to the individual (i.e., its nature, or status as a token of that kind, may be fixed by factors internal to the individual), even while that state is described, or identified (even by means of its content), by reference to factors external to the individual. I will argue (in section 5) that content ascription in Marr's theory is indeed an attempt to identify the content of a mental state, and hence identify the mental state, without purporting to supply the conditions in virtue of which that state is individuated. Marr leaves open entirely what properties of the individual or her environment determine that her state has the content that it does. Thus, while the individual's visual state may be identified by reference to the external environment, it could still be individuated individualistically.

2.3 Taxonomy

The individuation/identification distinction reveals how confusion can arise when the debate over individualism is framed as a debate about *taxonomy* in science. Egan (1992, p. 443), for example, begins her paper with the claim that "Individualism in psychology is a thesis about how mental states are to be taxonomized" (see also Egan, 1994, p. 258). And Burge (e.g., 1986) maintains that explanatory practice in the sciences, the classifications in terms of which the lawlike generalizations and regularities of scientific explanation are stated, should dictate to us how the kinds of that science are individuated. Taxonomy, however, need only be concerned with *identifying* kinds in ways that permit one to frame the lawlike (or counterfactual supporting) generalizations and regularities. As the sieve example illustrates, there may be generalizations that hold in virtue of the properties that individuate a kind, but that could nevertheless be framed in terms of non-essential properties that are only coextensive with the individuative properties. In the toy universe of discourse of the sieve example, subsuming the specific event of one ball passing through the device under a generalization stated in terms of the color of

balls will be just as effective an explanation as one stated in terms of the size of balls; no applicable instances will fail to be brought under that regularity if it is stated in terms of the color of balls. Nevertheless, as we saw, the regularity holds in virtue of the size of the balls, and hence the individuation of the kinds in this domain is in terms of the size of the balls, not their color. It is therefore not sufficient to determine whether kinds in a particular domain are individuated individualistically merely to look at how a theory or science of that domain taxonomizes states. The taxonomy may, without any loss of generality, be cast in terms of properties that do not individuate the kinds of that domain.

This point is important and worth emphasizing: Explanation in science is about showing how specific events are instances of lawlike regularities, or how specific objects conform to certain generalizations. Since these regularities and generalizations may be cast in terms of properties that do not individuate the events, states, or objects in question, one cannot blindly appeal to the classifications implicit in scientific explanations in order to individuate the kinds that fall within that science's domain. Yet this is the procedure that Burge (e.g., 1986) appears to advocate. I maintain that he is wrong to advocate this position. And when we see that individuation may, as it were, outrun the taxonomies implicit in scientific explanation, we see that we cannot simply look at the surface of scientific taxonomy to settle metaphysical questions about individuation. Taxonomies in science are guided by contingent restrictions on the types of experiments that can actually be run. Another example will help illustrate the point:

Suppose we want to test a device to see whether it is responding to triangles rather than trilaterals. If there are no trilaterals that are not triangles, then this is an experiment that we simply cannot run. Any regularities in the behavior of this device can be stated with equal generality in terms of trilaterals or triangles. Nevertheless, there may be a fact of the matter about what the device is responding to; it may be responding to angles rather than lines. If so, then there is a fact of the matter about how the device is individuated that is, as we might say, deeper than is revealed by the taxonomization of the device by the relevant science. We cannot distinguish between triangle-detectors and trilateral-detectors experimentally, so this distinction is not reflected in the taxonomy of the device. Nevertheless, the device might actually be a triangle-detector rather than a trilateral-detector, in which case a proper individuation of the device outruns the scientific taxonomization of the device.

As a matter of fact, however, virtually all theories do, to some level of refinement, taxonomize (or identify) states in virtue of their individuative properties. This is because individuation proceeds in

stages. Suppose there is a theory that quantifies over Fs; to be an F, the theory states, an entity must have properties x, y, and z. If the theory aspires to a further stage of individuation, it would say what must be the case for an entity to have properties x, y, and z (perhaps that requires having properties i, ii, and iii). Folk psychology, for example, individuates mental states in terms of attitudes and contents; a state must have an attitude component and a content component to be a folk psychological state. Folk psychology contains no explicit statement of what must be the case in order for a state to have the content that it does (that's why twin thought experiments were needed to argue for externalist claims about the individuation of folk psychological states). Thus, folk psychology individuates states to a certain level of refinement, but does not individuate states to the level of refinement that is at issue in the individualism debate.

When theories have explicit individuative commitments, I will call them *direct* individuative commitments; that is, when theories contain hypotheses about what properties tokens must have if they are to be of a certain kind, the theories have direct individuative implications with respect to those properties. Atomic theory, for example, does not merely identify gold by its atomic number, it holds explicitly that having atomic number 79 is an essential property of that kind. But many theories do not endeavor to individuate kinds down to a level that would settle a metaphysical dispute like the one involved in the individualism debate; that is, at certain stages of individuation, many theories are agnostic about what properties are essential to the kinds in their domains. Folk psychology, as we noted above, does not contain direct individuative commitments regarding content; that is, folk psychology does not say explicitly what facts must be true of a state if it is to have the content that it does (for if it did contain such explicit individuative commitments, the resolution to the individualism debate would have been obvious, and there would have been no need to resort to twin thought experiments (i.e., the T-I test) to settle the issue. To take another example, neuroscientists often identify brain areas by various techniques (ranging, e.g., from early staining methods to more recent imaging technologies) without antecedent views on the structural or functional properties that they later discover are key to individuating those areas (see Carlson, 1986, ch. 5).

2.4 *Direct and Indirect Individuative Implications*

As I will argue in section 4 (following Segal), Marr's theory does not have *direct individuative implications* with respect to content; by that I mean that Marr's theory does not specify the facts that make it true of a state V that it has content c. For that reason, it does not itself address the question of whether visual states are individuated

individualistically. But the theory may have *indirect individuative implications* with respect to content; by that I mean that it may have implications for the individuation of visual states by virtue of the application of the T-I test to Marr's theory. If Marr's theory assigns different contents to internally identical individuals in actual and counterfactual conditions that differ only in environmental factors, then those environmental factors would seem to be relevant to individuation; whether a state V has content c would in that case depend in part upon the environment.

Recall, however, that the issue of individualism is about how kinds are *correctly* individuated (see again, Burge, 1986, p. 3). I will argue that Marr's theory has indirect individuative implications of the sort just considered only if the relevant parts of the theory are true. Marr's theory may actually be false in certain details (an anonymous reviewer reminds me that many visual theorists believe that certain details are in fact false). If the theory is fundamentally misguided, or if the relevant parts of the theory are in fact false, then there would be no reason to accept any implications the theory may have for the individuation of visual states (in the same sense that there is no reason to accept the conclusion of an argument that is valid but not sound). Philosophical commentary on Marr's theory has uniformly assumed, however, that the theory is *not* fundamentally misguided, and that the relevant parts of Marr's theory are *not* false. The relevant parts of the theory include Marr's appeal to contents in individuating visual states, and the theory's ecological approach to content assignments. These are widely taken to be true, and (as I will discuss in section 7.2) they are presumed to be true for very good reasons.

Nevertheless, I will argue in section 7.2 that Marr's theory has no indirect individuative implications that bear on the issue of individualism, least of all any that favor anti-individualism. To foreshadow the discussion of section 7.2: A theory's indirect individuative implications, based on application of the T-I test, depend not just on whether the theory assigns contents correctly in the actual case, but also on whether the theory assigns contents correctly in the counterfactual case. I will argue that while there are reasons to think that the contents Marr's theory assigns in the actual case are correct, those reasons are not available in the counterfactual case. Thus, it is open to the individualist to maintain that Marr's theory would simply get the contents wrong in counterfactual cases. So, even though it may *attribute* different contents to twin individuals in different environments, there is no reason to expect that it would be right to do so, and hence no reason to think that the twins' visual contents really *would* be different. There would, then, be no reason to conclude from the

application of the T-I test to Marr's theory that visual states are correctly individuated non-individualistically.

3. *Burge's Argument*

In this section, I will review Marr's theory, and Burge's argument that it favors anti-individualism. I will then review Egan's and Segal's alternative interpretations of Marr's theory, and the response to Burge's argument that these interpretations suggest.

3.1 Marr On Vision

The structure of Marr's theory is well-known and bears only a minimal reminder here. The aim of the theory "is to understand vision completely, that is, to understand how descriptions of the world may efficiently and reliably be obtained from images of it." (Marr, 1982, p. 99.) Visual stimuli, like stimuli from other cognitive domains, underdetermine the representations they make possible in later stages of processing. Indeed, the processing in need of explanation is just the step-by-step augmentation of the contents of representations in the system. This explanatory enterprise devolves into three levels of theory. The computational level seeks to describe the functions that the visual system as a whole must compute, the tasks it must subserve. The algorithmic level seeks to provide the algorithms for the computation of the functions set out at the computational level of the theory; at this algorithmic level, the visual theorist must uncover the representational primitives of the theory, the information that must be encoded explicitly at the various stages of processing.[8] The implementational level of the theory describes the neural mechanisms on which the algorithm runs.

This analysis of the explanatory tasks facing the visual theorist may be somewhat oversimplified (see, e.g., Butler, 1994), but it does capture important aspects of explanation in cognitive science. Moreover, as almost all commentators have pointed out, Marr's analysis of the visual system, as well as other analyses carried out within the framework just mentioned (e.g., Ullman, 1979), have been quite successful in explaining the machinations of the visual system.[9] Indeed, this explanatory success will play a significant role in my dispute with Burge and Shapiro over the individuative implications of Marr's theory (see sec. 6).

3.2 Burge On Marr

Burge bases his analysis of Marr's theory on four examples of "how the theory treats the relation between the visual system and the physical environment". The examples he discusses (involving, e.g., zero-crossings and edge detection, stereopsis and distance detection, etc.)

have received considerable attention in the literature, so I will not rehearse them here. He draws several points out of these examples: First, he purports to show that in describing the processing of information in the visual system, Marr's theory "makes essential reference to the subject's distal stimuli and makes essential assumptions about contingent facts regarding the subject's physical environment." (Burge, 1986a, p. 29.) For example, in interpreting zero-crossing segments, Marr considers explicitly what in the distal environment such segments might correlate with, namely, real physical edges of some kind or another. Second, "the theory is set up to explain the reliability of a great variety of processes and sub-processes for acquiring information.... Reliability is presupposed in the formulations of the theory's basic questions." (Burge, 1986a, p. 29.) So, for example, the theory assumes that the visual system *solves* the problem of determining distance; that is, it is assumed that the visual system reliably generates accurate representations of objects as being at a particular remove from the subject. Third, Burge claims that in Marr's theory, "the information carried by representations -- their intentional content -- is individuated in terms of the specific distal causal antecedents in the physical world that the information is about and that the representations normally apply to." (Burge, 1986a, p. 32.) So, for example, Marr's notion of a generalized cone is derived from detailed consideration of the nature of physical objects with which the subject normally interacts.

As I suggested above, there is a last, more general, point about Marr's theory that Burge (and others) exploit: The theory has had significant predictive and explanatory success. That is, within the constraints of the theory, Marr is able to explain how the visual system could recover a 3-D representation of the visual scene. It is, moreover, able to explain how certain illusions are possible, and predict the circumstances under which they will occur. By virtually any measure, the theory is a success. This success functions as a reason to think that the theory is (at least approximately) *true*. And because it is true, there is reason to accept its individuative implications.

These points form the foundation for a much discussed, six-step argument that Burge offers to the conclusion that Marr's theory is not individualistic:

(1) The theory is intentional
(2) The intentional primitives of the theory and the information they carry are individuated by reference to contingently existing physical items or conditions by which they are normally caused and to which they normally apply.

(3) So if these physical conditions and, possibly, attendant physical laws were regularly different, the information conveyed to the subject and the intentional content of his or her visual representations would be different.

(4) It is not incoherent to conceive of relevantly different physical conditions and perhaps relevantly different (say, optical) laws regularly causing the same non-intentionally, individualistically individuated physical regularities in the subject's eyes and nervous system. It is enough if the differences are small; they need not be wholesale.

(5) In such a case (by (3)) the individual's visual representations would carry different information and have different representational content, though the person's whole non-intentional physical history might remain the same.

(6) Assuming that some perceptual states are identified in the theory in terms of their informational content, it follows that individualism is not true for the theory of vision. (From Burge, 1986a, p. 34.)

Step (1) is relatively unproblematic; Marr is clearly concerned to uncover, among other things, the contents of the representational primitives in early vision.[10] Steps 2 and 3 are supported by the general point, extracted from the examples Burge considers, that content is (purportedly) individuated by Marr's theory in terms of the normal distal cause of tokens of a given type of representation. It is this claim that Burge says "is tantamount to the chief point about representation or reference that generates [the] non-individualistic thought experiments" (Burge, 1986a, p. 27, fn. 14). Thus, Burge here takes Marr's theory to have what I have called *direct* individuative implications; there is no need to appeal to the T-I test to generate individuative implications, they are instead found in the theory's assignments of content. To reject (3), Burge claims, one would have to reject "the basic methods and questions" of the theory. Step (4), of course, cannot be denied by the individualist, and (5) follows from (3) and (4). Step (6) just reiterates (1), which together with (5) entails that Marr's theory is not individualistic.

3.3 Egan On Burge

In Egan's view, it is the first premise of Burge's argument that is mistaken. I think Egan's view is flawed fundamentally, and it will clear the way for, and illuminate, subsequent discussion if we expose the problems with her view before proceeding further.

Marr's theory is a computational theory, and computational theories, Egan claims, are concerned only with an articulation of the mechanisms of computation (see, e.g., Egan, 1992, p. 445). These mechanisms operate independently of any semantic or intentional properties, so a theory of the computational mechanisms of the visual system will have no use for any appeals to contents. She argues that all descriptions of computations in the theory can be given formal characterizations; descriptions in terms of contents are therefore superfluous (see Egan, 1992, pp. 453-455).

Contents may, however, figure in explanatory models of the theory, models that illustrate or help us understand what the theory says. She compares Marr's use of intentional terminology to other uses of explanatory models in the history of science:

> "Maxwell's efforts to represent Faraday's lines of force by the flow of liquid through tubes was an attempt to make intelligible a purely formal exposition of unfamiliar phenomena (in this case, electromagnetic phenomena) by appeal to systems governed by laws of mechanics, which have the status of familiar principles. In the case of mechanical models, the relevant similarity between model and modeled phenomena is a nomic isomorphism, that is, an isomorphism between two corresponding sets of laws. In the computational case, the interpretation function fI that pairs symbolic expressions with contents specifies an isomorphism between computational states and features of the represented domain." (Egan, 1992, p. 451.)

Egan points out that understanding the theory without appealing to contents would be difficult, and that the questions that define the psychological domain are usually couched in intentional terms. So, if we are to understand the theory, and be responsive to the relevant questions, we must appeal to contents, but only as explanatory models of the theory, not as part of the theory itself.

The way in which non-intentional computational theories explain the performance of cognitive tasks is, in Egan's view, analogous to the way in which positing an electrical field detector explains the ability of sand sharks to detect prey (see Egan, 1995). The shark detects prey *by* detecting electrical fields. Of course, the ability that the electrical field detector helps explain is not the ability to detect *prey* unless prey are (normally) detected by the shark; but that does not entail that the fact that animals in the sharks normal environment produce electrical fields is part of electromagnetic theory. Similar remarks apply in the case of computational explanations of cognitive processes. It is of course true that the process of recovering depth

information from information about disparity is not a visual process unless the states involved can be characterized in terms of information about depth and disparity; but that does not entail that the intentional characterizations are part of the computational theory that explains that visual process. Depth information is recovered from information about disparity *by* the computation of a formally specified function.

Because the theory itself has no essential use for contents, arguments to the conclusion that Marr's theory is not individualistic that appeal to the purportedly non-supervenient content of visual states cannot succeed. Whether the theory itself is individualistic depends on how computational states, *qua* computational, are individuated. And there is no doubt, according to Egan, that computational states, *qua* computational, are individuated without reference to the subject's environment.[11] Therefore, the theory is individualistic. Egan (1992, p. 444) grants that the contents assigned in perceptual psychology are generally non-supervenient, and hence that computational theories of perception, like Marr's theory, are often, in her words, 'externalist' about content. But because attributions of content are not essential in computational psychology, this content externalism leaves untouched the individualism of Marr's theory.

I disagree. Egan argues that the symbolic states and functions of computational theories have identity conditions independently of semantic interpretations. This is correct, but insufficient to make her case. No one disputes that Marr's theory is a computational theory. But that does not entail that it is *only* computational. That is, there is no justification for limiting the theory to a specification of the states and mechanisms in virtue of which it is computational. As a theory of *vision*, it is concerned with various aspects of vision, not just the computational mechanisms. In particular, it is concerned with the *contents* of representational primitives in the visual system. So, while a given *symbolic* state might have type-identity conditions that do not involve content, a given *visual* state, presumably, will not. It will be type-identifiable as a symbolic state, but also as a state with a particular visual content. There is no warrant for thinking that the theory offers only one sort of individuation condition for visual states; visual states have both computational and intentional identity conditions.

The point is easy to see when we recall the familiar fact that systems that are computationally identical may nonetheless be deployed in wholly different domains. One and the same program, individuated computationally (i.e., formally), may process visual and auditory information (see Davies, 1991, p. 482). Specifying the computational properties of a system, then, is insufficient for specifying whether it is a visual system or an auditory system. Part of what makes

a visual state a state of a *visual* system rather than an auditory system is that it has a *visual content* rather than an *auditory content*. It is not inconsistent with this claim to hold also that if a state is to be a state of the human visual system, then it must also satisfy a further condition cast in computational terms. But this does nothing to undermine the claim that it is part of the essence of a state of the visual system that it have a visual content. Egan's mistake is to assume that, because symbolic states have formal identities independently of contents, these formal properties are essential to the individuation of visual states while contents are somehow not essential to the individuation of visual states described in a computational theory of the visual system. But content assignments are not *optional* parts of a theory of vision. Of course, if the theory in question is a computational theory, as Marr's is, then specification of the computational properties of the states of the system will also not be optional. Both formal and intentional properties are, according to Marr's theory, relevant to the individuation of visual states.

This point undermines virtually everything Egan (1994) says in defense of her position against Morton (1993). For example, she says (1994, p. 259) that

> "...Marr typically characterizes [visual processes]...by reference to what they represent. ... The processes are also characterized *formally*...."

She then slides without further argument to this claim:

> "For the purpose of *individuation*, the formal characterization of a computational process carries the day."

Individuation *of what*? Visual processes, presumably; that, after all, is what Marr's theory is concerned with. But under this interpretation, Egan has committed a *non sequitur*. From the fact that visual processes have both an intentional and a formal characterization it does not follow that either one, let alone the formal, must "carry the day" in individuating *visual* processes. No doubt formal characterizations carry the day in individuating computational processes, but computational processes are not visual processes unless they process visual information. So, the fact that formal characterizations individuate computational processes is irrelevant to the question whether Marr's theory, a theory of *visual* systems, is individualistic. Egan makes the same mistake when she claims (1994, p. 263) that intentional descriptions of visual processes "...do not individuate the computational process." This, too, is true but irrelevant. A physical (or neural) process, according to Marr's

theory, must meet at least *two* conditions if it is to be a visual process: It must be a computational process, and it must be an intentional process (of a particular sort).

Given this, we can see that Egan's criticism of Morton's challenge is insufficient to block a more subtle challenge. Egan (1994, p. 260) alleges that Morton is committed to the claim that "a given computational state...could not have a different content and be a computational state of the same type". She then goes on to point out that this claim is false, so Morton's position is untenable. But if we replace "computational" with "visual", the claim becomes true, and it is Egan's position that becomes untenable.

Though Egan says, (1992, p. 452, fn. 16) that in her (1991, esp. pp. 198-202) she *argues* for the claim that content is merely "an adventitious feature of [the] theory's model", there is in fact only one short appeal to Marr's authority, where he claims of the primitives of the primal sketch that they are, as Egan (1991, p. 198) puts it, "...abstract properties of the retinal image." Two points mitigate the force Egan's observation. First, as Burge (1986a, p. 28, fn. 15) has noticed, there is some uncertainty about the precise stage at which the visual system, according to Marr's theory, "goes intentional". There is no corresponding uncertainty, however, that it does go intentional eventually. Second, even the passages from Marr that Egan quotes, and her interpretation of them, involve claims about the primitives of the primal sketch representing aspects of the retinal image. Thus, they are characterized intentionally even at this level. The only remaining question is where in the processing of this information the subsequent primitives come to stand for (i.e., represent) aspects of what the retinal image is an image of, rather than just the retinal image itself.

Egan (1994, p. 262) also appeals to Marr's authority and claims that

> "...construing the computational level as a mathematical specification of the function computed [rather than an intentional specification] explains Marr's insistence that the correct characterization of the algorithm depends upon the precise (read *mathematical*) specification of the function computed."

But this does nothing to bolster the claim that contents are not essential to the characterization of the function computed. Since visual processes are computational, according to Marr's theory, a correct specification of the algorithm does *depend upon* the mathematical specification. But the mathematical specification does not *exhaust* the specification of the algorithm. A complete specification of the algorithm would cast

the inputs and outputs in intentional terms that reveal the role of the algorithm in the processing of visual information. The appeal to content is in this case crucial. Thus, there is simply no plausible way to deny that, according to Marr's theory, the visual system is *essentially* a representational system, even while it is *also* essentially a computational system.

Egan's (1995) suggestion that Marr's theory explains visual processes in a fashion analogous to the way electromagnetic theory explains shark prey-detection is flawed fundamentally. An account of the detection procedure for the shark is given in terms of electromagnetic theory, the theory characterizing the nature of electromagnetism. In the case of vision, however, what plays the role analogous to electromagnetic theory is not Marr's theory, but *computational* theory, the theory characterizing the nature of computation; an account of, e.g., the recovery of depth from disparity is cast in terms of computational theory. Marr's theory is *a* computational theory, and thereby appeals to a theory of the nature of computation in identifying the computational stages involved in early vision. But that is hardly a reason to think that Marr's theory is *exhausted* by its appeal to computational theory; Marr's theory may *also* be an intentional theory. It is the application of computational principles to the visual system that distinguishes Marr's theory from the general theory of computation, and from non-computational theories of vision, just as it is the application of electromagnetic theory to shark prey-detection that distinguishes the electromagnetic theory of shark prey-detection from the general theory of the nature of electromagnetism, and from non-electromagnetic theories of shark prey-detection. Intentional characterizations are part of what makes Marr's theory the theory that it is.

Egan argues further that a non-intentional individuation of computational mechanisms is necessary to explain certain aspects of the biological fitness of the visual system. If the visual system is characterized intentionally, then in an environment that is systematically different, so that the representational primitives would be given systematically different intentional interpretations, the visual system would be of a relevantly different kind; we would not see "...why *this* visual system would *not* be adaptive had the environment been different." (1991, p. 201; see also 1994, p. 264) This argument attacks a straw position according to which Marr's theory, if it includes essential individuative appeals to intentional contents, eschews individuative appeals to the formal properties of computational mechanisms. As I argued above, however, Marr's theory takes *both* formal and intentional properties to be individuative. Thus, it is able to say how that same *computational* mechanisms might flourish (or not) in different

environments, even while it maintains that that computational mechanism in a relevantly different environment would be a different *visual* mechanism.

I note in passing, however, that the appeal to nonintentional properties is, in principle, not necessary in order to individuate a visual system in such a way that the same system could be considered in different environments. A visual system could be individuated intentionally in terms of the environment in which it evolved, and then considered (counterfactually or actually) in a relevantly different environment. In the new environment, it is still the system that evolved in the old environment, and the intentional content of the states of the system can still be characterized in those terms. An intentional specification of the functions computed by that system in the new environment would reveal that, while it is *facing* new information processing problems in the new environment, it is not *solving* them (*pace* Burge, 1986a, p. 35; see also section 7.2 below). Instead, it is still producing representations that are correctly characterized in terms of its old environment. It may still be adaptive, but this reveals only the familiar fact that adaptiveness does not require intentional states that represent the distal environment accurately (see, e.g., Segal, 1989, p. 208-9, and sections 5.2 and 7.2 below). Indeed, if intentional contents are determined individualistically, then this is precisely what we would expect to find. So far as I can see, there is no reason why Marr's theory could not accommodate individuation of this sort.

Egan also argues that her interpretation of computational theories as having the restricted goal of articulating only the computational mechanisms of the visual system (in non-intentional terms) is consistent with Marr's use of the intentional idiom at the computational level of his theory. The intentional idiom, she claims, is simply "...a recipe for achieving the correct characterization of cognitive mechanisms" (Egan, 1992, p. 445). I grant that this heuristic interpretation of the computational level of the theory is consistent with the goal of the theory being restricted to explaining computational mechanisms. But, apart from the antecedent implausibly of her position (see above), this interpretation of Marr's use of the intentional idiom at the computational level is certainly not the interpretation of Marr that is most faithful to his words. Marr claims explicitly, for example, that he aims

> "...to understand vision completely, that is, to understand how descriptions of the world may efficiently and reliably be obtained from images of it. ... What kind of information does the human visual system [encode]...? How does it [encode] this information...?" (Marr [6], p. 99)

Marr gives every indication here and elsewhere that an attempt to characterize visual states independently of their informational content is seriously misguided; a "complete" understanding of the visual system cannot neglect its representational nature. Thus, while it is conceivable that Marr's use of the intentional idiom serves only a heuristic purpose, there is nothing in Marr's discussion that motivates such an interpretation.

Egan's case for the claim that appeals to content are located in explanatory models of computational theories, and not in the theories themselves, now comes down only to the analogy with Maxwell's model (cf. a similar analogy at 1991, p. 200, with Bohr's theory of the atom), and the claim that contents will help in understanding the theory and answering the questions the theory was devised to answer. Unfortunately, the analogy fails. Whereas tubes and liquids are not part of the phenomenon that Faraday's appeals to forces are intended to explain, intentional contents, we have noted, *are* part of the phenomenon that Marr's theory is intended to explain. The visual system is a cognitive system, and any theory that purports to understand vision completely had better make reference to the very feature that makes the visual system a cognitive system, namely, visual contents. Moreover, it is not an accident that the questions the theory was set up to answer are cast in the intentional idiom. Contents are part of the phenomena about which we have questions, indeed, the phenomena the theory is supposed to explain.

For the most part Egan only *suggests* her interpretation of Marr's theory in her various writings on the subject; very little *argument* for the claim can be found. I suspect that Egan's use of examples such as an adding machine, and her incorporation of Cummins' interpretational semantics (1992, p, 450-1), lead her to view contents as non-essential properties of the visual states described by Marr's theory. But visual states, unlike states of an adding machine, are inherently intentional. Egan's (1995) suggestions about the explanatory role of content, while potentially informative, also fail to appreciate the essential role of content in visual state individuation; content assignments are not made true by the extent to which they connect computational processes to their normal environments. Rather, the intentional facts about visual states, whatever may fix them as they are, determine the truth-value of content assignments. Content *may* connect computational properties to normal environments, but only if the system represents its environment accurately. Egan's attempts to explain away Marr's use of the intentional idiom do nothing to cast doubt on this claim (and even taken as they are intended, as attempts to alibi Marr's talk of content, they are unconvincing). We must bear in mind that Marr's theory, while it is

a computational theory, is also a theory of vision. It may individuate *computational* states individualistically, but that does nothing to suggest that it individuates *visual* states individualistically.[12]

3.4 Segal On Burge

Segal argues that Burge's analysis of Marr's theory is flawed. I will present here a brief synopsis of Segal's reply so that the stage is set for the subsequent presentation of Shapiro's commentary on the Burge/Segal dispute. I will therefore present Segal's case in roughly the way Shapiro sees it. In section 4, we will take a closer look at Segal's reply, and extract an important point that Shapiro has apparently overlooked.

Segal bases his assessment of Burge's argument on careful consideration of the constraints on content assignments in Marr's theory. He argues that every assignment of content must meet three conditions:

First, there must be a bottom-up account of how a representation with a given content could have been derived reliably from the previous stage in visual processing. Since visual stimuli underdetermine later visual representations, the contents encoded at various stages of visual processing must contain more information than was available in the contents of representations at the previous stages. Segal's first constraint is that there must be no stage at which the content of the representation cannot be accounted for by appealing to the representations that precede it in processing, and the activity of the processing mechanisms.

Second, each content assignment in the theory must respect a top-down constraint. We have just seen that in order for there to be a possible bottom-up account of a particular stage of processing, the representational content being augmented must not radically underdetermine the representational content to which gives rise causally. Therefore, in determining how much augmentation occurs at any given stage, the theorist must look to the later stages of processing to make sure that the content assigned makes the later contents possible. These determinations will both help constrain, and be constrained by, the choice of properties of the physical instantiation of the representations that are encoding the contents. As I understand Segal's suggestion, the visual theorists' judgment here may be informed by various global assessments of the tasks the visual system is performing.

Third, there must be some evidence that the visual system is actually employing the representation assigned to it. The conditions so far mentioned underdetermine the theorist's choice of content assignment, so, to rule out at least some of the candidate assignments that meet the previous two conditions, each content assignment must be checked against behavioral evidence, particularly, evidence of the

subject's discriminatory capacities; the theorist should not assign contents that are more fine-grained than the subject's discriminatory capacities allow. So, if the system is said to be able to distinguish an O from a C, there must be evidence of this in the system's observable behavior.

Armed with these conditions, Segal considers a "twin" example of Burge's. In this example, certain of P's early visual representations are regularly caused by shadows; for this reason, Burge claims, P's percept represents a shadow. If this percept were occasionally caused by an indistinguishable crack, then P would misrepresent the crack as a shadow. In the counterfactual case, the situation is reversed; certain of Twin P's early visual representations are regularly caused by cracks, and for this reason represent cracks. If this percept were on occasion caused by a shadow, Burge claims, Twin P would misrepresent the shadow as a crack.

Segal claims that Burge has adopted an overly simple view of content assignment according to Marr's theory. Burge assumes that the content of a representation is given by what in the distal environment normally causes a representation of that type. But mere causal covariance, Segal argues, does not provide adequate insight into the contents the visual system is employing. In assigning the contents that he does, Burge is assuming, without evident warrant, a causal theory of reference and representation; the content of the representation is given by the essence of the kind of distal object or event that normally causes a representation of that type. Segal calls the assignments made on such a theory *restrictive*; the twins would be assigned contents, shadow and crack respectively, that are specified in such a way that the twins have different contents in spite of their internal identity. According to Segal, the three conditions mentioned above, especially the third, merit more *liberal* content assignments, on which the twins' visual states have the same content; each twin represents the thin, dark lines in their distal environments as, Segal claims, *crackdows*. Thus, when it comes to the T-I test, the twins do not have different contents. So, there is no indirect anti-individualistic implication stemming from Marr's theory.

4. Shapiro's Argument

Shapiro (1993) maintains that, while Burge's argument that Marr's theory is non-individualistic fails, the theory still has direct individuative implications that favor anti-individualism. Moreover, he claims, this conclusion follows even on Segal's own analysis of Marr's theory. On this basis, he then offers the surprising contention that the T-I test is in fact not diagnostic of individualism. In this section, I will review Shapiro's diagnoses of Burge's argument and Segal's reply, his

(Shapiro's) assessment of the T-I test, as well as his (Shapiro's) own analysis of Marr's theory.

4.1 Shapiro On Burge and Segal

In Shapiro's view, Burge's six-step argument fails spectacularly. It is advertised as a carefully motivated argument against individualism, but in Shapiro's view, it merely begs the question against the individualist at (2). Recall that (2) claims that contents on Marr's theory are individuated by appeal to "contingently existing physical items...". If this is indeed the case, if the theory does indeed have direct individuative implications, then there is no need to proceed further with careful reasoning that deploys the T-I test; the non-individualistic implications of Marr's content assignments can be gotten straight from the content assignments, which appeal to contingently existing features in the subject's environment. In going forth with steps (3)-(6), Burge is merely demonstrating the uncontroversial fact that if a theory is not individualistic, it will fail the T-I test. The argument is therefore "not the sophisticated piece of reasoning [Burge] would have us believe", and simply begs the question at step (2).

Shapiro claims that the direct individuative implications of Marr's theory also serve to obviate Segal's reply to Burge. When Segal argues that Burge's restrictive content assignments are not justified by Marr's theory, he does not question the basic *source* of those assignments, namely, the covariation with elements of the distal environment. What Segal offers is another way of individuating the distal environment, so that the twins' visual states end up being assigned the same contents by the theory, in which case the theory passes the T-I test. But, since it is still the distal environment that governs the content assignments, the theory cannot be individualistic; it appeals to the environment in individuating the contents. Shapiro concludes: "Segal, far from arguing that Marr's theory is individualistic, seems to depend upon the theory's non-individualism to argue that it passes the T-I test!" (Shapiro, 1993, p. 496).

Moreover, Shapiro claims, Segal has no warrant to claim that discriminative behavior is a source of content assignments on Marr's theory. If the computational theory mandates a description of the distal environment that favors Burge's restrictive content assignments, then it will override any constraints issuing from discriminative behavior. Thus, according to Shapiro, there is no generally applicable liberal interpretation of Marr's content assignments; as we will see in a moment, whether the contents of twins visual states will be of identical types is determined by the computational level, and there is no guarantee that it will favor liberal (type-identical) contents.

4.2 Shapiro On Marr and the T-I Test

Shapiro believes that Marr's theory does indeed have direct individuative implications: Contents are assigned, on the theory, by appeal to the normal distal cause of a given representation-type, therefore the theory is non-individualistic.[13] But he does not think the contents are assigned in the simplistic manner Burge assumes, nor in quite the way Segal suggests. Rather, he claims, the computational level of the theory, the level at which the task of the visual system is described, plays the dominant role in determining content assignments. According to Shapiro, Marr determines the task of the visual system by appealing to the evolutionary purpose of vision. Once this task is determined, it is used by the theorist to constrain the description under which a representation's normal distal cause is represented. As Shapiro puts it,

> "It is *because* the task of the visual system is to provide accurate and reliable descriptions of surfaces in the world that the representations serving as inputs to the algorithms that perform this task must stand for features of surfaces in the world." (Shapiro, 1993, p. 506.)

For example, the task can be accomplished only if (or at least if) certain changes in the intensity of the gray array constitute the representation of an edge. Thus, what content (description) the theory assigns to a particular state of the visual system is dictated by the computational level of the theory, particularly, what in the world the visual system would have to represent if it is to fulfill the function for which it is evolved (see Shapiro, 1993, pp. 498-508).

It is this essential appeal in the visual task description to contingent features of the distal environment that, in Shapiro's view, assures us that Marr's theory, as it involves content assignments, is not individualistic. Yet, Shapiro asserts, in spite of the theory's non-individualism, it would, under the right task description, *pass* the T-I test. That is, if the task of the visual system is described in such a way that, say, Segal's liberal content assignments are the ones dictated by the computational theory, then twins will be assigned the same contents, and hence pass the T-I test. So, since a non-individualistic theory can nevertheless pass the T-I test, the T-I test cannot be diagnostic of individualism. The explanatory and predictive success of Marr's theory vindicates both the theory, and this verdict with respect to the T-I test (see Shapiro, 1993, pp. 508-510).

Thus, while Burge and Segal do disagree about whether content assignments on Marr's theory will be liberal or restrictive, and hence

whether the theory will pass the T-I test, their disagreement is, according to Shapiro, independent of the issue of individualism.

5. *Epistemic and Metaphysical Projects*

I will argue in this section that Shapiro and Burge, unlike Segal, fail to appreciate the essentially epistemic nature of Marr's approach to content assignments. Once this error is made apparent, we will be able to see how Burge's argument fails (though not in quite the way Shapiro claims), and how Shapiro's own argument misses its mark.

5.1 *The Distinction*

The problem that content assignments pose for Marr's theory is essentially an *epistemological* problem. By that I mean to make this claim: The aim of the theory is to discover, or figure out, what contents certain visual states (i.e., the representational primitives of the visual system) actually have. I do not dispute that, in determining what contents to assign, Marr makes reference (e.g., through the task analysis) to contingent features of the distal environment, and the subject's individual and evolutionary history of interactions with that environment. But the environment, as well as evolutionary speculations, serve as heuristics that help Marr figure out what contents to assign, and what descriptions are most helpful in assigning those contents.

Content assignment, for Marr, does not pose a *metaphysical* problem. By that I mean to make this claim: Individuation is a metaphysical issue; the environmental and evolutionary factors to which Marr appeals in the task analysis at the computational level of the theory, however, do not fix, or individuate, the states to which the contents are being assigned; they do not tell us what properties the visual system must have in order to have representational primitives with those contents. It is worth noting as well, in conjunction with this claim, that the descriptions of the assigned content are referentially opaque. In describing the content of representational primitives as, say, "edge", Marr is not thereby claiming that real edges in the distal environment constitute part of the supervenience base for states with that content; nevertheless, the states are about edges.[14] It is in this sense that Marr's theory does not have direct individuative implications; the theory does not tell us what facts must be true of a token state if it is to have the content that it does. It should not be assumed that Marr's content assignments take environmental features to participate in the supervenience base of visual contents.[15]

5.2 *Computationalism and the Epistemic/Metaphysical Distinction*

The epistemic/metaphysical distinction is, I will assume, clear enough. I will now argue that the epistemic interpretation is the

preferred interpretation of Marr's discussions of content. Shapiro seems to miss or disregard the distinction even though Segal is quite explicit in deploying the distinction in his discussion of Burge and Marr. Segal (1989, p. 199) says:

> "Burge claims that Marr uses [elements of the distal environment] to *individuate* contents. In my view, he merely uses them as a guide to discovering what the contents are. So, to pick a neutral expression: the items and conditions [in the distal environment] are those in terms of which Marr *identifies* contents." (Emphasis added.)

Segal supports his claim that Marr's approach to content assignment is, as I am calling it, epistemically oriented rather than metaphysical, by considering several quotations from Marr that either favor the epistemic reading, or at least fail to favor the metaphysical reading of the theory (see Segal, 1989, pp. 211-213).[16]

Indeed, consideration of Marr's overall project belies a metaphysical interpretation of his approach to content assignments. As we noted above, Marr aims to *understand* how the visual system operates. There is no place in the book, or even in any of the subsequent developments of the computational approach in the psychological literature, where attention is directed at the metaphysical issue of content individuation. He never raises the issue of the supervenience base for contents, and never discusses what properties individuals must have in order to have states with a particular content. All that is claimed is that certain environmental correlations are required if the theory is to assign the content that it assigns.

To see that this claim is independent of any claim about individuation that would settle the individualism debate, look at the difference between the following two questions.

A. What facts must be true of a state if *the theory is to assign* content c?

B. What facts must be true of a state if *it is to have* content c?

A and B are different questions, and can quite easily receive different answers even if the theory in question is true. Suppose that a theory appeals to facts that are correlated with those facts that are relevant to individuation, but that are not themselves individuative. To take the example from section 2.3 above, we might have a theory that classifies devices by their detection of sides of a figure, even when the

device in fact detects angles. This classification will preserve all the regularities and generalizations that would be preserved under an angle classification provided that there is an appropriate correlation between angles and sides.

In the case of Marr's theory specifically, we know that the methodological constraints on content assignment by the theory are such that the theorist is directed to look at certain environmental features in assigning contents. Thus, the answer to question A will include facts about the environment. But this in no way guarantees that the answer to question B must include facts about the environment. For suppose that, unbeknownst to most current researchers on vision, there is some internal, physical property of visual states (call them wavelets) that is responsible for the intentional properties of visual states; suppose, that is, that the supervenience base for visual contents is some local physical property, in which case individualism is true. And suppose also that evolution has selected visual systems such that these systems generally deploy locally supervenient representations of Xs when Xs are present in the visual field; thus, the veridicality assumption of Marr's theory turns out to be true. In this case, Marr's methodology for content assignment will in fact assign the correct contents; it will *assign* to a visual state the content X when the state in fact *has* the content X. But this will be because Marr's methodology for content assignment exploits a (contingent) correlation between the environmental features that visual states represent, and the local physical properties that determine (metaphysically) the contents of visual states; that is, there will be a correlation between the objects of visual representation and the supervenience base for visual representations. Thus, Marr's content assignment methodology may be perfectly legitimate, and assign to states the contents that they actually have, even if the facts the theory appeals to in assigning contents are not the facts that determine metaphysically the contents that the visual states actually have. Thus, while the answer to question A may include facts about the environment, it is perfectly conceivable, consistent with the success of Marr's theory, that the answer to question B does not include facts about the environment.

This entire discussion could be carried out in the context of other contributors to computational vision theory. In citing the work of Ullman (e.g., 1979), an anonymous reviewer has suggested that in many cases, the functions that are computed by visual systems must be characterized in terms of real physical properties in the environment. Ullman says, for example, that:

> "...Meaningful units are...symbols in the representation whose meanings are founded in the environment, not in the intensity array. (Ullman, 1979, p. 12)"

Ullman argues further that the inputs of certain functions (in this case the so-called correspondence function) cannot be facts about the intensity array of the image of the retina because

> "...the changes of the raw intensity distributions do not directly reflect changes in the visible environment. Hence, the organization of the visual input into units corresponding to physical entities is a prerequisite for the recovery of physical motion from the changing optical array. (Ullman, 1979, p. 18).

And in discussing the Cornsweet illusion (where a visible edge can have radically different intensity distributions), Ullman again shows that the function computed must take features of the distal environment as inputs. For differences in illumination and orientation of the same physical edge can give rise to different intensity distributions that "are of no interest to the perceiver who is to recover the physical structure of the environment".

These examples all illustrate that, according to computational theories of vision, at least some of the functions that the visual system computes take representations of objects or properties external to the individual as inputs. But the point I have been pressing is that this characterization of the functions computed by the visual system does not answer the question of individuation of relevance to the individualism debate. The question of what function the visual system is computing is, according to computational theories of vision, answered by appealing to the representational contents of the states of the visual system; and I am agreeing that these states represent features of the distal environment. But individualism does not deny that states of the visual system represent objects or properties external to the individual. Individualism denies only that the facts in virtue of which states of the visual system represent these distal objects and properties are facts external to the visual system (it also denies that any historical properties of the system are relevant to individuation by representational capacities). Thus, even though Ullman tells us that the functions computed by the visual system cannot be stated in terms of representations of the intensity distribution of the retinal image, that does not begin to call for a verdict in the dispute over individualism.

Returning to Marr's contribution to computational vision theory, Shapiro's main claim that, according to Marr himself, the computational level of the theory is primary even in the metaphysical

arena of content individuation, misconstrues Marr's intent. When Marr says, for example, that trying to understand the algorithmic and implementational levels without understanding the computational task of the system is like "trying to understand bird flight by studying only the feathers", he is giving expression to an epistemic point. The aim is one of discovery, not individuation.

As we discussed above in section 2.3, scientific theories in general need not make claims about individuation. Thus, if Shapiro and others are to claim that Marr's theory makes claims about individuation that would settle the debate over individualism, it is incumbent upon them to *argue* that Marr's theory should be interpreted to be making such claims. In the absence of such an argument, there is no reason whatever to construe Marr's approach to content assignment as having anything at all to say about what properties a visual state must have if it is to have the content that it does.

The strongest reading needed is one on which Marr states that the theorist must use the environmental evidence if she is ever to determine what contents visual states actually have. We may even grant, as we did above in the context of Ullman's contribution and as we should in any case, that the representations that help define what is being computed are about features in the environment. But this interpretation is absolutely silent on the question of what facts about the world make it the case that the representational states have the contents that they do. Hence, it is also absolutely silent on the question of what makes it the case that the functions that are computed are as they are. For the computations derive their identity from the representations that constitute their inputs and outputs, and whether the representations derive their identity from factors internal to the individual or not is precisely the question that is up for grabs. Indeed, it is entirely consistent with Marr's methodology for content assignments and Ullman's methodology for computation identification that visual states could have the contents they do even if individuals never interacted with the environment; if that were the case, of course, the visual theorist would have no way of discovering what the contents are until individuals interact with the environment long enough for regularities to emerge. But that is in no way inconsistent with the methodologies under discussion. As a matter of contingent fact, it may be that it is through interaction with the environment that the visual system comes to have the states that it does; but that causal story of the genesis of visual states with distal content again entails absolutely nothing about what *individuates* the states.

The clinching point seems to me to be this: Suppose again that it turned out that visual states (and other neural states that are candidates for being psychological states) have some special physical

property that is responsible for the intentional properties of the states in question. That is, suppose that there is some unique physical property of neural states in virtue of which they represent the properties of the environment that they represent. So described, this would constitute an entirely internal supervenience base for the contents of visual states. Thus, the properties in virtue of which visual states have the contents that they do are internal properties; individualism is true. Now ask whether this would show that Marr's method of assigning contents is wrong? Would this expose a flaw in Ullman's claim that the computations of the visual system take representations of the distal environment as input?

Obviously not! Even given all of this, it may still be sound methodology for assigning contents, and therefore discovering (and stating) what functions are being computed; it may, indeed, be the only or best methodology for assigning contents or characterizing computations. For we may have no other way of figuring out what the contents are; the physical properties that individuate the states with those contents, the properties in virtue of which the states have those contents, may be too difficult to discern or detect. In this case, it would make perfect sense to use Marr's methodology to assign contents, or Ullman's methodology for discovering and describing contents. The environment still provides the best *evidence* of what the contents are, and the contents still provide the characterization of what is being computed. It is just that the environment would not *individuate* the states with those contents, nor then would it individuate the processes in virtue of which functions from one content to another are computed. So, because it is entirely consistent with Marr's theory and Ullman's theory that individualism is true, it cannot be a direct commitment of these theories that individualism is false. Thus, there is no argument for anti-individualism that appeals only to the content and computation assignment methodology of computational vision theory. Philosophers who read computational theories as having direct anti-individualistic implications simply misread them; they import into the theory metaphysical commitments that there is just no reason to believe are there.

Thus, it is not an accident that Shapiro must provide an argument for his metaphysical construal of Marr, rather than a quotation in which Marr reveals his individuative intentions. In discussing content assignments on Marr's theory, Shapiro cites the content Marr assigns to zero-crossing segments, and then gives his (Shapiro's) own reason for thinking that the content is fixed by the task of the visual system:

> "...if it were not the purpose of the visual system to provide descriptions of surfaces in the world, then the fact that the physical world constrains the interpretation of the zero-crossings in the primal sketch would not matter. Why look to the world to see how its structure constrains the structure of the primal sketch if one does not view the goal of visual system to be the production of a reliable description of the world?" (Shapiro, 1993, p. 508).

Shapiro's argument, however, establishes a conclusion quite different from the conclusion he wishes to draw from it. As Shapiro puts it, the physical world does not constrain the contents of the zero-crossings, but only constrains the "interpretations" of them. Interpretations, though, are not contents, but hypotheses theorists' make about contents. It is the visual theorist's hypotheses about what contents the zero-crossings have antecedently (prior to her investigation) that the physical world constrains, not the contents themselves. Similarly, it is the visual theorist who must "look to the world" to figure out how the world constrains, not the structure of the primal sketch, but the visual theorist's own hypotheses about the structure of the primal sketch. In both cases, the constraints at issue are epistemic constraints on what hypotheses to accept, not metaphysical constraints that help fix the content (and hence type-identity) of visual states.

Many of the quotations that Shapiro himself considers betray the epistemic slant of Marr's theory. For example, in discussing the role that evidence of discriminative capacities should play in discovering representations and algorithms, Shapiro (p. 501) considers the following quotation from Marr.

> "First, there is usually a wide choice of representation. Second, the choice of algorithm often depends rather critically on the particular representation that is employed. And third, even for a fixed representation, there are often several possible algorithms for carrying out the same process. Which one is chosen will depend on any particularly desirable or undesirable characteristics that the algorithm may have..." (Marr, 1982, p. 26).

Marr's use of words such as "choice", "depend", and "desirable" deserve scrutiny. He is obviously not claiming that the visual theorist must choose which factors *to use* to fix (metaphysically) the kinds of representations in the visual system; visual theorists do not play *any* role in fixing anything about the system they are studying since they neither design nor build it. Marr is claiming only that the visual

theorist must choose from among candidate hypotheses about the representations and algorithms that the visual system actually uses. The theorist will base her choice from among those candidates on certain theoretical desiderata. Thus, it is the *choice* that depends on the desirability of the characteristics, not the representations or algorithms themselves. In other words, there is nothing of metaphysical or individuative import going on here, no discussion of what properties an individual must have in order to have states with a given content. Marr is concerned entirely with the epistemic goal of discovering which representations and algorithms are to be found in the system.

Shapiro appears to be sensitive to this fact when he compares a designer's *choices* to the visual theorist's *hypotheses*. If our goal is to design, say, a cash register, then which choices we make about algorithms will constrain our choice about how to represent numbers. However, if our goal is to understand an extant system, such as the visual system, then our hypotheses about algorithms will constrain our hypotheses about how features of the distal environment are represented. In the design case, our choice will play the equivalent of a metaphysical role in determining or fixing the representations and algorithms. In the theoretical case, however, our choice has no such influence over the system; it is merely part of an epistemic (or discovery-oriented) project.

To his detriment, however, when Shapiro adapts this discussion to the case of content assignment, he ignores the very distinction just mentioned:

> "...behavioral evidence provides data about which algorithm and so which representations underlie some task, but this evidence only helps to limit choices among representations the contents of which are constrained and fixed by the computational level of the theory." (Shapiro, 1993, p. 503.)

It is plain that the "*choices* among representations" are in fact *hypotheses* about which representations to postulate; and these hypotheses, Shapiro argues, are constrained by the computational level of the theory, as well as by discriminative behavior. The content assignments, too, are constrained by the computational level of the theory. But for some reason, Shapiro is given to believe that the constraint here is not on *hypotheses* about content, i.e., content assignments, but on the very contents themselves. As we have seen, though, there is no warrant whatever in Marr's discussion to make this jump when talking about content or any other part of the theory.

It is likely that Shapiro is misled by Marr's cash register example, and by his own example of an artificial representational system. The artificial system involves bottle caps being used by someone to represent either the number of beverages consumed, or the innings played in a baseball game. Obviously, in this example, the contents of the bottle caps are fixed by an outside interpreter, rather than by factors in the system of bottle caps (individualistically) or by, in addition, its domain of representation (non-individualistically). This should not surprise us since, of course, the bottle cap system is representational only in a metaphorical sense (or at least a derivative sense), and quite unlike the sense in which the visual system is representational. The contents of the bottle caps are derived from another representational system, the interpreter, while the contents of the interpreter's mental states, and of course the contents of the visual system with which Marr is concerned, are, as Dretske would put it, *intrinsic* (which is not yet to say internal) to the visual system. The contents of bottle caps may be determined by the goals of the person using the bottle caps because the bottle caps do not have their contents intrinsically. But the contents of visual states are not like that; their contents *are* intrinsic, and have nothing whatever to do with the goals of some outside interpreter. So, what goes for bottle caps should in no way be assumed to go for the visual system.

Thus, what Shapiro should have said is not that the task of the visual system somehow makes the representations stand for features of the distal environment, but only that the task, as *discovered* from considering evolutionary factors, plays a heuristic role in the *identification* or description of the contents of the representational primitives of the visual system; this leaves it entirely open how such contents are fixed or individuated metaphysically.

In spite of the considerations I have so far presented, though, Shapiro may yet want to claim that Marr's theory may be, as it were, "deconstructed" to reveal a metaphysical view of content fixation and individuation. He may think, in other words, that the process by which the visual theorist figures out what contents to assign recapitulates the process by which contents are fixed metaphysically. But quite aside from the lack of warrant for interpreting Marr in this way, there is an inscrutable indeterminacy that plagues metaphysical views of this sort:

The appeal to evolution by natural selection as a mechanism for fixing the task description of the visual system is insufficient. Broadly speaking, the evolutionary function of any system or organ within an individual is to assist the individual in surviving to reproductive age, and eventually reproducing. But if this is the function of the visual system, then it is possible that it can fulfill that function without generating reliable representations of the distal environment. For

example, an organism that represents only a disjunction of retinal states that correlates perfectly with features of the distal environment, rather than the features of the distal environment, will suffer no loss of fitness. That is, the visual system might represent only the retina, and leave it to Nature to make the states of the retina correlate with factors in the distal environment. So, if the evolutionary function of the visual system is specified only broadly in this way, there is no motivation for the specific task description on which the visual system would generate representations with distal contents.

In biology, however, organs are often attributed more finely individuated functions on the basis of evidence about the specific way in which the organ contributes to the overall fitness of the organism in which it is embedded. Perhaps we can do the same with the visual system, and fasten on a description of its task that requires it to generate representations of the distal environment, and even particular features of the distal environment. In the case of a representational system like the visual system, however, there is no way for us to observe the system to determine what more specific contribution it might be making; that is, there is no evolutionary reason appeal to which will help us decide between different sorts of contribution the visual system might make. So, there is no way to determine what content assignments (e.g., distal vs. retinal) are required for the system to make its contribution to the fitness of the organism. Thus, the appeal to evolutionary functions as a way of fixing contents metaphysically shows little promise of success. We do, of course, take the visual system to have a more fine-grained function, and hence distally specified contents, but as yet we have no theory to account for these functions and their attendant contents.

Though full discussion of this issue must be left for another occasion, it is worth pointing out that attempts by Millikan (e.g., 1984, 1991) and Shapiro (1992) to show how selectional contexts might exhibit the requisite opacity and intensionality are insufficient. We are left inevitably with the recalcitrant problem that mechanisms can be selected for their ability to represent properties co-extensive with the adaptively relevant property. Appeal to evolutionary pressures may provide us with *evidence* that adaptively relevant properties are in fact represented, but they are insufficient to *individuate* contents finely enough. Dennett (e.g., 1987, p. 311ff), Fodor (e.g., 1987, ch. 4), and Davidson (in conversation) develop this criticism in ways that neither Millikan nor Shapiro fully come to grips with.[17]

This metaphysical view of content fixation is, moreover, question-begging against the individualist, since it appeals to factors other than an internal supervenience base in fixing contents metaphysically. Therefore, we should expect that the individualist reaction to Shapiro's suggestion, construed in this way, is simply to reject

it. Given that this metaphysical interpretation is entirely irrelevant to the successes of Marr's theory, and unmotivated by Marr's text, this seems indeed to be the proper response.

Given the remarks of this section, we are now in a position to see that Egan was right at least to notice that Marr's remarks about content are *not* concerned with the *individuation* of visual states.[18] Recall that Burge (1986a, p. 3) tells us that "Individualism is a view about how kinds are correctly individuated, how their natures are fixed." I have argued that intentional properties play an essential role (even if not a sufficient role) in individuating visual states into kinds; they are part of the nature of visual states. So, whether or not the theory is individualistic depends, in part, on whether the theory maintains that the content of a visual state is fixed, or determined to be what it is, by factors internal to the individual, or in part by external factors. The fact that the contents of visual states (and cognitive states more generally) are expressed by Marr in terms that describe the distal environment says nothing about what it is that makes it the case that a particular visual state has the content that it does. The content attributions describe what the visual states represent, not what it is in virtue of which they have that representational content. Appeals to specific contingent features of the individual's environment are helpful in discovering what contents visual states actually have, but this does not entail that these aspects of the distal environment are relevant to the individuation of visual states. There is no suggestion here that the contents of visual states supervene wholly or partly on the environment. About this much Egan was right.

In my view, however, her interpretation goes too far in that direction. She thinks that the very appeal to content is a heuristic that aids us in discovering computational mechanisms and in understanding a theory that harbors no essential commitment to contents. My claim has been that while Marr is a realist about the contents of visual states, his approach to content assignment concerns only how investigators might figure out what contents to assign to states of the visual system. Thus, it is my claim that Marr offers heuristics for the assignment of *particular* contents, but that there is no reason to think that the assignment of contents *at all* is itself a heuristic move.

5.3 Consequences for Shapiro's Argument

For Shapiro's main argument, the consequences of having missed the metaphysics/epistemology distinction are unfriendly. First, the assumption that Marr's theory has direct individuative implications is revealed to be false. Marr uses assumptions about the task of the visual system, and descriptions about the normal distal causes of representations, in forming hypotheses about the contents of

representational primitives in the visual system. But these facts entail exactly nothing about how the contents are fixed metaphysically, and individuated into kinds. Specifically, it is entirely consistent with Marr's methodology that the contents he postulates are fixed solely by factors internal to the individual or the individual's visual system.

Given this, we can see that Shapiro's claim that the T-I test is not diagnostic of individualism is completely unsupported. That claim, recall, depended on the claim that Marr's theory has direct individuative implications. Shapiro then claimed that the theory would, under certain circumstances, individuate contents widely and still pass the T-I test by assigning the same content to twins in different environments. Since we know now that Marr's theory does not have direct individuative implications, there is no way to determine, prior to the application of the T-I test, whether it individuates contents widely. It may identify contents by reference to environmental factors, but that entails nothing about individuation. Therefore, there is no way to show that the Marr's theory would individuate contents widely yet still pass the T-I test. As far as Shapiro has shown, whether the theory individuates contents widely can *only* be determined by application of the T-I test.

For exactly the same reasons, Shapiro has no right to claim that Segal has conceded anti-individualism in arguing for his liberal interpretation of content assignments on Marr's theory. Segal claims that Marr is identifying contents by appeal to the distal environment, not individuating them. Shapiro's failure to notice the epistemic nature of the problem of content assignment is responsible for his failure to see the distinction Segal is making here. This, in turn, is responsible for Shapiro's mistaken view that Marr's theory has direct individuative implications, and, subsequently, his erroneous criticism of Segal. If Segal's view is mistaken, it is not mistaken for the reasons Shapiro provides.[19]

5.4 Burge and the Epistemic/Metaphysical Distinction

Shapiro claims also that Burge merely *assumes* that Marr's theory is not individualistic at step (2) in the vaunted six-step argument that Burge advertises as *proving* that conclusion. Shapiro is right to point out that Burge's argument is, at the very least, suspicious at just this point. Burge does say in step (2) that

> "the intentional primitives of the theory and the information they carry are *individuated* by reference to the contingently existing physical items or conditions by which they are normally caused and to which they normally apply." (Emphasis added.)

If Burge really does mean to say that visual states are *individuated* in this way, then he really has begged the question. But we should be suspicious of an interpretation on which Burge's main argument is so manifestly question-begging. Burge exploits a twin thought experiment in this context, I suggest, because it is needed to show (or to attempt to show) that visual states are individuated non-individualistically. As it turns out, I think that Burge is frequently incautious in his use of the term "individuate", and there is some reason to think that occasionally he uses the term to mean something like *identify*, in the sense set out in section 2.3.[20] Given this, it is certainly more plausible to proceed under the interpretation that Burge is merely expressing in step (2) the claim that the theory identifies or assigns contents in terms of (or by appeal to) the normal distal causes of the states that have that content.

In addition, we must bear in mind that the theory is a theory *of* the visual system, and the states in question are states of the *visual system*, not states in the *theory*. Therefore, when Burge refers in (2) to "the intentional primitives of the *theory*", we should interpret him to be referring to the intentional primitives of the *visual system*, as it is described in the theory.

In light of these suggestions, we can rewrite steps (2) and (3) of Burge's argument in ways that make the argument more plausible (adjustments are enclosed in brackets []):

> (2') The intentional primitives of the [visual system (according to the theory)] and the information they carry are [identified or assigned] by reference to the contingently existing physical items or conditions by which they are normally caused and to which they normally apply.
>
> (3') So, if these physical conditions and, possibly, attendant physical laws were regularly different, the information conveyed to the subject [according to the theory] and the intentional content of his or her visual representations [as assigned by the theory] would be different.

Steps (2') and (3') make Burge's argument essentially an application of the T-I test. Step (4) effectively asks us to imagine a counterfactual situation in which the environment is relevantly different, and step (5) tells us that the theory would assign different contents to the counterfactual visual system. Step (6) can then be interpreted to be asserting that, according to Marr's theory, the contents of the

representational primitives in the visual system are not individuated individualistically. Whether or not this is the argument Burge had in mind, on this construal of Burge's argument, he does not beg the question by assuming that Marr's theory has direct individuative implications. Rather, by applying the T-I test, Burge is claiming that Marr's theory has indirect individuative implications that favor an anti-individualistic view of the visual system.

Notice that my suggested interpretation of Burge's argument differs significantly from that offered by Shapiro (1993, pp. 493-4). Just as we have done here, Shapiro imagines an interpretation of the argument where Burge entertains the application of Marr's theory to actual and counterfactual visual systems, and claims that Marr's theory would assign different contents to the states of the two visual systems. Shapiro then repeats his charge that Burge has begged the question against the individualist by assuming that Marr's theory (directly) individuates contents by appeal to contingent features of the distal environment. But here Shapiro is constrained by his own misreading of Marr. Shapiro assumes that Marr's theory has direct individuative implications. I am suggesting an interpretation of Burge's argument on which Marr's theory has no direct individuative implications, but on which it does have *indirect* individuative implications, through application of the T-I test; these implications favor anti-individualism. That is, if the theory merely identifies or assigns contents by appeal to distal features, then it takes application of the T-I test to determine whether it entails that visual states are individuated non-individualistically. If it assigns different contents in the actual and counterfactual cases, then the theory's individuative implications will favor anti-individualism; if not, not.

With Shapiro's misreading of Marr out of the way, we can entertain an argument more like the one that Burge may have intended to offer. In the next section, I will revisit Segal's reply to Burge's argument (under the revised interpretation). I will argue that individualists will need to do more to stop Burge's argument. In section 7, I will argue that Burge's revised argument can be resisted even without Segal's liberal interpretation.

6. Segal's Liberal Strategy Reconsidered

Segal's liberal strategy, recall, recommends assigning the same contents to the twins in Burge's example involving shadows and cracks; they both represent crackdows. We saw above (section 4.3) that this strategy survives Shapiro's initial attempt to criticize it. But Davies (1991) raises two broad sorts of objection (one of which was originally Burge's) to which Segal has not provided adequate responses.

6.1 Liberal Perceptual Content and Restrictive Conceptual Content

Segal argues that content assignments on Marr's theory must conform to top-down constraints. This condition requires the theorist assigning content at a particular stage of processing to look to the contents assigned (or to be assigned) at later stages, and to make sure that the content assigned at the earlier stage makes possible a bottom-up account of the content assigned at the later stage.

If we extend this condition, so that it applies to the contents assigned at the highest stage of visual processing, then it will command the theorist assigning these higher-level contents to look to the other sorts of (non-visual) representations that that content has to make possible. This extended top-down constraint is derived from an overall view of the role of visual processing in the larger cognitive system of which it is a part. Thus, we can accept Shapiro's point that content assignments are derived in part by appeal to the computational level of the theory, provided we export from it the metaphysical reading that he prefers.[21] Indeed, I think Shapiro shows quite effectively that Marr does look to the computational level of the theory to help in the assignment of content.

What sort of lessons are learned when we consider the larger role of visual processing? Davies (1991, p. 481) points out that the "non-conceptual" outputs of visual processing constitute the inputs to conceptualized thoughts and judgments. Burge (1986a, p. 43), too, notes that "...vision provides intentional input for other cognitive capacities..." (see also Burge, 1986a, p. 25-6). Thus, Davies (1991, p. 481) argues,

> "...to grasp a concept that is at least partly observational -- say, the concept of a cube -- a subject must have a disposition to judge that an object is a cube if the object is perceptually presented *as* a cube. And -- avoiding circularity -- this latter notion is cashed out in terms of the non-conceptual content of the [visual] experience of the object being that there is a cube spatially related in such-and-such a way to the subject.
>
> This theoretically desirable close mesh between concepts and non-conceptual [i.e., visual] content is not possible if there are no experiences in which an object is presented as a cube. And there will be no such experiences once [Segal's liberal strategy] has been applied."

Thus, if we take the conceptualized contents of thought and judgment to be providing top-down constraints on visual representations, we will have to posit visual contents that make a bottom-up account of conceptual content possible. And, because conceptual content is

unquestionably restrictive, this constraint suggests that we will need, not the liberal content assignments that Segal favors, but the restrictive ones that Burge favors.

Segal (1989, p. 203-4) seems to be aware of the possibility of this sort of top-down constraint, but concludes nevertheless that the assignment of contents at the level of thought and judgment (i.e., at the level of folk psychology) does not constrain the assignment of contents at the level of visual representation. In fact, he agrees that Burge may well be right to adopt restrictive assignments of thought contents at the level of language and folk psychology. But, because sciences often fall out of conformity with common sense, and because they must determine their own individuative criteria, Segal is not troubled.

However, given Segal's own endorsement of top-down constraints in lower-level content assignments, it is hard to see how he can maintain consistently that top-down constraints cease to be relevant when they comefrom consideration of later processing.[22] He admits (1989, p. 208) that he cannot rule out the possibility that some such constraint will motivate restrictive content assignments, but he does not entertain any such constraint. I believe that the considerations raised here and by Davies (and, perhaps also by Shapiro) do offer a plausible suggestion. Segal (1989, p. 205 and p. 206-7) does express confidence that Marr would favor the liberal reading, that the attribution of restrictive content to visual states would be "out of line with Marr's actual practice", but there is no argument in the surrounding discussion that even bears on that conclusion. Instead, Segal argues that restrictive contents would be unmotivated under certain conditions. But he fails to consider the visual system in its larger context, and so is not responsive to challenge posed here.

Segal (1991, p. 491) does offer a direct response to the challenge as Davies (1991) posed it. He says that

> "[V]isual *cubes* feed into cube judgments. *Crackdow* representations could lead into perceptual judgments that go: *there's one of those things again, that could be, say, a crack or a shadow.*"

But this response misses its mark in two ways. First, the challenge that Davies poses involves a twin cube story; so that challenge is to say how we arrive at judgments with the content *cube* when our discriminative capacities do not allow us to distinguish cubes from twin cubes. Merely *saying* that cube representations in the visual system will feed into cube judgments is not responsive to this challenge; Segal must tell us how such cube representations are possible in the context of a twin story.

Segal does argue earlier that a twin cube story is not likely to be forthcoming. His reason for thinking this is that he can imagine all sorts of ways that the twin's behaviors might diverge if they are assigned representations with different contents (e.g., *cube* and *twin cube*). For example, when asked to trace with their hands what shape they are seeing, unless they have visual representations with the same contents, their behaviors will diverge. This is what Segal calls his first VIP (very important point): It is not plausible that there is an anti-individualistic twin story on which both twins interact successfully with their environments.

But this VIP is not adequate. Individualism fails if there are *any* circumstances under which internally identical subjects have different contents. Suppose the difference between the twins' representations (on restrictive assignment) is due in part to suitably different laws of optics and physics; in that case, an object with a different shape would be, in that environment, indistinguishable from a cube, as viewed in this environment. Moreover, the twins would trace the same shape when they are tested in the same environment, but different shapes when tested in different environments. So, it is not enough to claim that the appropriate twin story is implausible; after all, *every* twin story is implausible. The real test, the only test, is whether they are possible. The twin cube story surely is.

Second, Davies complaint is not that the liberal contents will not lead to *any* perceptual judgments, but whether they will lead to the *right* perceptual judgments. Thus, Segal's crackdow representations could lead to judgments with the contents he supposes, but the facts of the story are that we, possessors of human visual systems, do not make those judgments; we judge the lines before us to be whatever normally causes them. That, in fact, is how illusions are possible; the same visual data are consistent with different distal interpretations, but, as Burge's restrictive interpretation of Marr's theory supposes, the contents of our representations are in agreement with the normal distal cause, not all indistinguishable distal causes. If Segal is going to admit that Burge is right to assign restrictive contents in the case of thoughts and judgments, then he owes us an account of how judgments with *those* contents are possible consequences of visual processing, not how judgments with any content at all are possible. As far as I can see, he has not explained what needs explaining.

6.2 *Objectivity*

This last point dovetails with the second main objection to Segal's liberal strategy. Burge (1986a, pp. 39-43; see also Burge, 1986b) offers an argument designed to supplement the anti-individualistic

interpretation of Marr's theory. The argument can be stated in the form of three premises.

(i) Our concept of objectivity is such that no one objective type that we visually represent is such that it must vary with, or be typed so as necessarily to match exactly, an individual's proximal stimuli and discriminative abilities.

(ii) Some objective physical objects and properties are visually represented as such; they are specifically specified.

(iii) We individuate intentional visual representations in terms of the objective entities that they normally apply to, for members of a given species.

Given these premises, Burge argues, individualism must fail.

I will not pause here to critique this argument.[23] Our concern is only with the first premise, which articulates a conception of objectivity with which analyses of perceptual representation must engage. This notion of objectivity, as I mentioned above, is needed to account for illusions (what Burge calls "fundamental misperception"). In my view, Segal's liberal strategy never does engage this notion of objectivity adequately. He claims (1991, p. 491), that his first VIP (mentioned above) allows him to account for objectivity of this sort. But we have seen that this VIP is inadequate. The fact that the appropriate twin cases are implausible is of no help to the liberal strategy; it still cannot explain how one can represent some object as a cube, given that there are other possible shapes that might cause a representation of that (e.g., syntactic or formal) type. Therefore, Segal has no right to claim that on his liberal interpretation of content assignments, "a subject might fundamentally misperceive something as a cube." As far as he has shown, the liberal interpretation will assign contents with extensions that include everything that is indistinguishable from the normal distal cause restrictively individuated, provided only that the subject's fitness in her environment is not seriously threatened.[24]

Given that the liberal strategy does not obviously allow for the application of a concept of objectivity like that sketched in the first premise of Burge's supplementary argument, and given that it does not conform well to the restrictive content assignments of higher-level cognitive systems (and folk psychology), it would be desirable, from the point of view of the individualist, to have some other strategy with which to resist Burge's anti-individualistic interpretation of Marr's

theory. In the next and final section of this chapter, I will offer such a strategy.

7. *The Conservative Defense Strategy*

A conservative individualist with respect to Marr's theory claims that, even on restrictive interpretation, visual states are correctly individuated individualistically. Davies (1991, p. 466-468) considers a defense of conservative individualism that, while it fails, is close to a strategy that I think does succeed. I will begin by considering the failed strategy and its shortcomings. I will then turn to the successful one.

7.1 An Unsuccessful Defense of Conservative Individualism

We saw earlier (section 4.3), in contrast to Shapiro's claims, that Marr's theory could be revealed to have indirect individuative implications by applying the T-I test to the theory to see what verdicts it yields in actual and counterfactual cases. Consider Burge's shadow/crack example. If Marr's theory would assign the content *shadow* to P's visual state, and *crack* to Twin P's, even though they differ only in their normal environment, the individuative implications of Marr's theory would be revealed to favor anti-individualism. Given the results of the last section, this might seem to be inevitable.

But, as Davies (1991, p. 466) points out, there may be a way out for the conservative individualist. It is a basic assumption of Marr's theory that "the gray level arrays seriously underdetermine the nature of the distal stimulus". The visual system must, through various stages of processing, build up to descriptions of the distal stimuli. We may assume that evolution has selected a system that does this effectively. In spite of this, Davies (1991, pp. 466-7) suggests,

> "It is open to us to suppose that the creature is doomed to be the victim of a good deal of misrepresentation. This verdict on an imagined counterfactual case is in no way inconsistent with "the success-orientation of [Marr's] theory" ([Burge, 1986a], p. 34), because success-orientation governs the methodology of the theory as it is applied to creatures in the actual world. If we now imagine one of those creatures, or a duplicate, in a quite different and inhospitable environment -- an environment to which the creature's visual processing system is not adapted -- then the methodology does not require that, even in those counterfactual circumstances, success should be the general rule."

Marr's theory assumes that our representations are fairly reliable, such that the normal causes of tokens of a given representational type are a

reliable guide to the content in terms of which those tokens are individuated into types. Marr assumes that the primary cause of this accuracy in visual representations is evolution; we could not get along in our environment well if we did not represent it in a reliably accurate fashion. Davies is suggesting that if that evolutionary cause is not present, if in the counterfactual situation the visual system is not well-adapted to its environment, then there will be no reason to think that its representations are accurate. So, if there is no evolutionary pressure that drives the visual system to token accurate representations of the distal environment, there will be no reason for thinking that Marr's method will yield accurate content assignments.

We can crystallize the point Davies is driving toward in the following way. To get even indirect individuative implications out of Marr's theory, we would have to satisfy four conditions:

(a) The theory has to yield a determinate content assignment in the actual case.
(b) There must be justification for thinking that this content assignment is correct.
(c) The theory has to yield a determinate content assignment in the counterfactual case.
(d) There must be justification for thinking that this content assignment is correct.

In the case under consideration (a) and (c) are clearly satisfied; Marr's theory assigns *shadow* to the state of the actual visual system, and *crack* to the state of the counterfactual visual system. (b) is also satisfied. But (d) is not satisfied in the case described. No evolutionary pressure has been exerted on the counterfactual visual system, so there is no reason to think that its contents are reliably accurate. But for this reason, there is also no reason to think that Marr's approach to content assignments is applicable in this type of situation. Thus, this sort of example is inappropriate as a test of the indirect individuative implications of Marr's theory.

7.2 A Successful Defense of Conservative Individualism

Though it was easy to spot the flaw in the original defense of conservative individualism, I believe that the general strategy of the defense can be resuscitated so that it is no longer vulnerable to the charge Davies levies against it. Ultimately, I will argue, there is no way to satisfy condition (d) even when the counterfactual visual system is well-adapted to its environment.

It is widely assumed that the contents Marr's theory assigns to the actual visual system are accurate. By this I mean to say that while

the content assignments aim only to identify the contents of visual states, they do so by describing the types of things in the world that the visual states really do represent (even if they have those contents in virtue of some feature internal to the visual system). There are two sorts of reasons we might consider to justify the assumption that Marr's content assignments are accurate.

First, as we saw above, these assignments seem more likely to be accurate given the (causal) evolutionary pressure in favor of reliable representational systems. But evolution cannot guarantee that a well adapted system will always token accurate representations; fundamental misperception, for example, must be possible, and routine mistakes are frequent. Moreover, evolution cannot even guarantee that representations are frequently or reliably accurate. Evolutionary pressure may *seem* to make it more likely that the success-orientation assumption of Marr's theory is true, but it cannot guarantee its truth. Evolution is too blunt an instrument to secure accuracy of contents possessed of intensionality that is more fine-grained than the extensions of the contents. Any given creature may get along in its environment making systematic errors in representation, so long as those errors are of the right sort. As Segal (1989, pp. 208-9) puts it, "...evolution only selects for behavioral success: creatures that do the right thing for the wrong reason do not have a tendency to die out" (see also section 6.2 above). So suppose, for example, that I represent all cracks as shadows, and all shadows as cracks. Provided that this error is of no evolutionary consequence,or that I invert my behavioral reactions to these items (so that responses appropriate to cracks are triggered by shadow representations, etc.), I will be well enough adapted to my environment.

Burge (1986a, p. 36) is aware of this fact:

> "In principle, we can conceive of some regular variation in the distal causes of perceptual impressions with no variation in the person's individualistically specified physical processes, even while conceiving the person as well adapted to the relevant environment -- *though, of course, not uniquely adapted.*" (Emphasis added.)[25]

There is, then, nothing in the appeal to evolution that would prevent either an actual individual, P, or a counterfactual individual, Twin P, from being, as Davies put it, "the victim of a good deal of misrepresentation", while all the while being well-adapted to their environments. If this is indeed the case, then Davies' sole reason for rejecting the original defense of conservative individualism is unavailable in the present context. Specifically, he has no reason with which to resist the individualists' contention that condition (d) above is

not satisfied; no individuative implications of Marr's theory would then be forthcoming, particularly no individuative implications that favor anti-individualism. Burge's argument would therefore be unsuccessful.

There is, however, a second aspect of Marr's theory that, while it does not suggest anything further that would help to guarantee *causally* that we have reliably veridical visual systems, might nonetheless vindicate *epistemically* the hypothesis that we do have such reliably accurate systems for visual representation. As I noted earlier, much of the interest in Marr's theory is due to its success in explaining certain visual phenomena (e.g., the generation of 3-D representations, illusions, etc.), and predicting the circumstances under which those phenomena will arise. Success of this sort is generally regarded as prime evidence of a theory's truth. So, one reason to think that Marr's success-oriented content assignments are correct is the explanatory and predictive success of the theory.

But, as a preliminary point, notice that explanatory and predictive success provides no assurance of reliable accuracy in visual representation, and hence no assurance that content assignments that assume as much are correct; abductive inference is not demonstrative. So, it could be that content assignments not sanctioned by Marr's success-orientation are actually correct. That is, in spite of the predictive and explanatory success of Marr's theory, the success-assumption of the theory, that our visual representations are for the most part veridical, might be mistaken; in that case, the contents assigned by the theory would not be the contents our visual representations actually possess. For example, it could be, for all we can tell or show, that our visual representations actually have twin contents; suppose that what we think are cracks are actually shadows and *vice versa*. We could then tell a coherent, but wholly different, story of the visual system's machinations. Thus, the assumption that the theory's content assignments are accurate is not entailed by the theory's predictive and explanatory success.

In considering Marr's theory in application to twin cases, however, we assume that the contents it assigns to the actual visual system are identified correctly. And, in spite of the fact that the correctness of the assignments is not assured, let us grant that we are right to assume that the contents Marr's theory assigns are the contents our visual representations actually have. The preliminary point I want to make, however, is that, because there is no guarantee that our visual systems reliably represent the world accurately, there is no guarantee that Marr's success-oriented content assignments are accurate content assignments. So, even though (I am granting) these assignments are correct, the accuracy of the identifications is not fully justified by the reasons we have considered.

Now consider the assignments the theory makes to Twin P's counterfactual visual system. As we noted above, there is no reason why Twin P cannot be well-adapted to his environment, yet still be the victim of regular and systematic error in his visual representations. This shows that the stipulation in the original defense of conservative individualism (i.e., the stipulation of a counterfactual creature that is not adapted to its environment) is irrelevant to the conservative defense strategy; the stipulation can simply be removed in favor of the stipulation of a creature that is well-adapted to its environment, yet still the victim of regular and systematic representational error. For this reason, the assignments of content that Marr's theory makes to Twin P's counterfactual visual system *could* be wrong.

This shows that Burge's way of applying Marr's theory to twin cases is not sufficient to *establish* the anti-individualistic conclusion; it is possible (though still just barely possible) that the contents that Marr's theory would assign in the counterfactual case are wrong. And given the possibility that they are wrong, we are not obligated to accept that they are right. So, we are not obligated to accept, as the anti-individualist argues, that Marr's theory *shows* that the twins will have different visual contents in spite of their internal identity; the theory *assigns* different contents, but we are not obliged to accept that the theory is right to do so. And without that, there is no argument for anti-individualism.

The anti-individualist is likely to respond to this, however, by saying that, while the counterfactual content assignments *could* be wrong, there is as much reason as in the actual case to think that they are *not* wrong. So, there remains substantial, if slightly weakened, support for the claim that the twins would have different visual contents, and hence that the anti-individualistic argument goes through after all.

But this is false. In the first place, evolutionary pressures, even in the counterfactual case, are consistent with inaccurate content assignments in the counterfactual case. Marr's theory may simply exploit the fact, if it is one, that evolution has (actually) selected visual systems that, e.g., deploy individualistically determined representations of X when Xs are present in the environment (i.e., visual systems that satisfy the so-called success assumption of the theory). It is consistent with this to suppose that evolution might have (counterfactually) selected visual systems that do not deploy representations of X when Xs are present (it might, e.g., deploy representations of twin Xs that are equally adaptive, and thus fail to satisfy the success assumption in a way that has no evolutionary consequences).

In the second place, the explanatory and predictive success of the theory is inapplicable in the counterfactual case. Thus, while the theory has had explanatory and predictive success in application to the actual visual system, there has been no matching explanatory and predictive success with respect to the counterfactual visual system of Twin P. There is nothing illegitimate in imagining a counterfactual situation in which Twin P's visual system is well-adapted, but it would beg the question against the conservative individualist simply to assume that Marr's theory would enjoy explanatory and predictive successin application to this counterfactual visual system. For suppose that conservative individualism is true; in that case, Twin P's visual contents would *match* P's, and if Marr's theory would assign different contents to the twins, the theory would be wrong in the counterfactual case, even if it would still be right in the actual case. To deny this without argument is to assume the very point at issue.[26] The theory has no track record at all with respect to the counterfactual system, much less one of explanatory success. There is, then, no reason at all to suppose that the contents Marr's theory assigns to Twin P are correct.

Thus, we now have a legitimate argument that condition (d) on the generation of individuative implications is not satisfied; there is no justification for thinking that (and no way of knowing whether) the theory's counterfactual content assignments are correct; it would beg the question simply to assume so. Without such justification, there is no way to establish even indirect individuative implications of Marr's theory, much less individuative implications that favor anti-individualism.

So, far from having, as Davies (1991, p. 468) says, "limited room for maneuver" in response to Burge's argument, the conservative individualist is able to block the argument definitively. Specifically, it is step (3) in Burge's argument that is flawed. Burge (1986a, p. 35) says, in support of step (3):

> "If the regular, law-like relations between perception and the environment were different, the visual system would be *solving* different information-processing problems..." (emphasis added).

The conservative defense strategy reveals this claim to be question-begging. There is no warrant for supposing that the counterfactual visual system would be *solving* a different problem, even if it would be *facing* a different problem. To deny that it would be solving this different problem does not require one to deny "the basic methods and questions" of Marr's theory; it is fully consistent with the theory's success-orientation, and its explanatory and predictive success, in the actual case, that it assigns contents mistakenly in the counterfactual

case, even in the counterfactual case where the twin is well-adapted. Indeed, there is no more reason in the counterfactual case to think that the theory would assign contents correctly than incorrectly. Because the anti-individualist cannot support the content assignments of Marr's theory in the counterfactual case, she has no hope at all of unseating the conservative individualist by appealing to twin cases.

The conservative defense strategy does not show that Marr's theory is "individualistic"; indeed, it grants that the theory would assign different contents to twins. But this is irrelevant to the issue of individualism, which concerns the individuation, the correct individuation, of visual states and processes. The conservative defense strategy therefore denies that the fact that Marr's theory would assign different contents to twins has any implications at all for the individuation of visual states, even while it grants the impressive support for Marr's theory in application to actual visual systems. Neither Burge nor Davies (nor Shapiro) has established that visual states, even when they are individuated by appeal to content, are individuated non-individualistically. Rather, consideration of Marr's theory in application to twin cases shows nothing at all about the individuation of visual states; it therefore shows nothing at all that could favor anti-individualism.

8. *Concluding Remarks*

Burge opens the closing section of his (1986a) this way:

> "Although the theory of vision is in various ways special, I see no reason why its non-individualistic methods will not find analogs in other parts of psychology. In fact, ...since vision provides the intentional input for other cognitive capacities, there is reason to think that the methods of the theory of vision are presupposed by other parts of psychology."

Given the absence of any reason to embrace Burge's anti-individualism in the case of vision, these remarks suggest that there will be a corresponding absence of reason to embrace anti-individualism anywhere in a computational psychology, indeed in any scientific treatment of psychology. The remaining chapters of this book are intended to bear this out.

CHAPTER 5

CAUSAL POWERS

1. *Preliminary Remarks*

If the results of the previous chapter are to be believed, then one influential way of arguing for anti-individualism in a scientific context, the appeal to computational vision theory, does no more than beg the question against the individualist. While this is, if true, an important result, it does little to motivate the search for a plausible form of individualism. In the context of the propositional attitudes of folk psychology, we found that certain epistemological pressures can be exerted on the externalist thesis, and these were taken to provide at least a *prima facie* motivation to develop a plausible internalistic account of the attitudes. I think we can derive a similar motivation in the scientific context by considering certain metaphysical pressures that can be exerted on the anti-individualistic thesis. This metaphysical pressure comes from considerations having to do with the causal efficacy of intentional states. I will argue in this chapter that the only plausible account of mental causation is inconsistent with an anti-individualism about the mental. That is, if there is any sense in which intentional properties are to be part of the causal order, they cannot be relational in the way that anti-individualism would have it.

This issue arises in the context of science insofar as it has been argued on both sides that a proper appeal to science will settle whether anti-individualism is required to countenance causal powers that are inconsistent with basic scientific commitments concerning causation and causal powers. Thus, although individualism is a *metaphysical* thesis, Fodor (1987, ch. 2, 1991) has argued that Burge's anti-individualistic individuation of mental states is incompatible with general constraints on individuation in science having to do with causation and causal powers. According to Fodor, it may be that common psychological discourse is permitted to individuate mental states externally; scientific psychology, however, cannot since it must, like other sciences, individuate states by appeal to causal powers, and these are limited to the internal states of an individual.

Several philosophers have countered this move.[1] I think they have shown that central aspects of Fodor's case are problematic. I will argue in this chapter, however, that these philosophers fail to see how

scientific explanations (and in particular the explanations of a scientific psychology) bear on the issue of metaphysical individuation. A proper interpretation of these explanations, together with the need for an account of mental causation, does indeed put pressure on anti-individualism in a way that motivates the search for a plausible individualism.

We will join the debate by considering the argument of Fodor's that others have addressed; this argument will be outlined in section 2. Though the argument might deviate in certain respects from one Fodor would be willing to embrace these days, it is this argument that Burge addresses, and so it is this argument that allows us entry into the debate. I will call it a Fodorian argument because it is clearly inspired by arguments Fodor has himself offered. This argument appeals to two key claims:

(STCP) scientific taxonomy is by causal powers,

and

(CPN) causal powers are restricted to the internal properties of an individual (i.e., they are narrow, in this sense).

It is CPN that Fodor seems less than sanguine about. Nevertheless, there are several important ways in which CPN can be defended. STCP, however, is quite problematic, and it will be one of the main points of this chapter that STCP should be rejected. I will therefore offer a variation on Fodor's argument that does not appeal to STCP.

I will focus on two broad types of rebuttal that Burge offers of the Fodorian argument. The first accepts STCP and argues from that premise to the rejection of CPN. I will call this the argument from scientific taxonomy. The second attempts to remove one sort of defense of CPN (from Fodor, 1991), which builds on the context test (CT), the idea that sameness of causal powers should be assessed across contexts; here Burge appeals to differences in the behavior of twins in the thought experiments. I will call this the argument from behavior. Neither of these arguments, I will claim, is effective against the revised Fodorian argument that I will present.

2. *A Prima Facie Case Against Anti-Individualism*

In this section, I will lay out a *prima facie* case against anti-individualism. This case begins with the Fodorian argument that Burge addresses. But several metaphysical points emerging from the previous chapter call for a modification of that argument. Both the Fodorian

argument and the repaired version of it appeal to certain claims about mental causation; sections 2.3 and 2.4 contain the presentation and defense of a certain picture of mental causation that, I will claim, is inconsistent with anti-individualism. In sections 3 and 4 below I will defend my variation of the Fodorian argument against the sorts of objections anti-individualists might raise against it.

2.1 A Fodorian Argument

Fodor's argument against anti-individualism about content appeals to general claims about taxonomy and causality in science. Burge has summarized the Fodorian argument as follows:

> (1) Psychological explanation is causal explanation.
> (2) States and processes appealed to in psychological explanation should be type individuated in terms of their causal powers.
> (3) The anti-individualistic conception postulates the possibility of differences in mental states and processes between two individuals without any corresponding differences in their brain states (internally individuated).
> (4) But individuals cannot differ in their causal powers if they do not differ in their brain states (internally individuated).
> (5) Psychological explanation should not type individuate states and processes anti-individualistically. (From Burge, 1989a, p. 305.)[2]

Premise (2) contains the commitment to STCP, and (4) contains the commitment to CPN. Burge's main interest is in premise (4), and hence CPN. In taking issue with (4), Burge is claiming that internally identical individuals (or individual brains) may nonetheless differ with respect to their causal powers; their causal powers, in other words, may be *wide*. If the causal powers of brain states, *qua* mental states, can be relationally individuated in the sense that anti-individualism takes content to be relationally individuated, then the anti-individualist will be at no comparative disadvantage with respect to the individualist when it comes to explaining mental causation. I intend to show that no case for such wide causal powers is forthcoming. Burge will therefore be unable to resist the Fodorian argument at step (4).[3]

2.2 A Repaired Version of the Fodorian Argument

Burge's case against (4) leans heavily on accepting (2). One of my main criticisms of Burge will be that (2) is, strictly speaking, false,

so any criticism that rests on (2) will be similarly undermined. For similar reasons, I claim that the conclusion of the Fodorian argument is ever so slightly ill-posed. This whole debate, recall, is about a metaphysical claim concerning the individuation of psychological states. I will be exploiting and developing a point raised initially in the previous chapter that distinguishes between scientific taxonomy and metaphysical individuation. Science contributes to the project of individuation, but that project is not exhausted by the contribution of science. In other words, metaphysical individuation cannot simply be equated with taxonomy in science. In important respects, individuation goes deeper than taxonomy. Thus, a taxonomy in the science of psychology can be stated in terms that appeal to an individual's relations to her environment even though the correct metaphysical individuation of psychological states requires no similar appeal to relational properties.

One way to view the relationship between taxonomy in science and individuation in metaphysics is that taxonomy simply *is* individuation of the sort that can be discovered (or confirmed, or supported) empirically. Scientific taxonomy and metaphysical individuation are both after the same goal: The discovery of Nature's causal joints. But science is limited to empirical procedures, while metaphysics avails itself of the results of experiments entertained in thought rather than just those performed in the lab. It may be the opinion of those with an empirical bent that thought experiments are insufficiently constrained to yield results deserving of our trust. I disagree; thought experiments are constrained by logical principles (e.g., consistency) and antecedent metaphysical commitments, most (if not all) of which are shared by our scientific brethren. It may be unwise to place *as much* trust in the results of thought experiments, but unless we are willing to forego all metaphysical claims and commitments, we had better be willing to place some stock in this tool of metaphysics.[4] Ultimately, then, it is the combined effort of science and metaphysics that will uncover the causal order of the world.

In light of this point, both (2) and (5) in the Fodorian argument must be recast in terms that reflect this distinction. Thus, the argument I will actually defend runs as follows.

(1) Explanations of psychological events and behavior is causal explanation.

(2) States and processes appealed to in psychological explanation may be taxonomized in terms of properties other than their causal powers, but the metaphysical individuations of those states and processes must cast in terms of causal powers.

(3) The anti-individualist individuation of mental states postulates the possibility of differences in mental states and processes between two individuals without any corresponding differences in their brain states (internally individuated).

(4) But individuals cannot differ in their causal powers if they do not differ in their brain states (internally individuated).

(5) Psychological explanation may taxonomize states and processes by appeal to relational properties, but the metaphysical individuation of those states must not be anti-individualistic.

I will argue that it is this conclusion that the debate between individualism and anti-individualism is about, and that Burge has done nothing to defend anti-individualism against an argument of this sort.

2.3 Mental Causation, Supervenience, and Tropes,

The repaired version of the Fodorian argument, like its predecessor, begins with the assumption that psychological explanation is causal explanation. There is an implicit hope in this assumption that psychological states participate in the causal relations that are the subjects of these explanations *in virtue of* their psychological properties. S goes to the refrigerator to get some ice cream *because* she believes that there is ice cream in it, and she wants some ice cream. Note that it is because S believes *that* and wants *this* that she does what she does; it is the intentional properties of her states that are supposed to explain (causally) why she undertakes ice cream gathering behavior. So, since intentional content is essential to the psychological identity of the states appealed to in these explanations, the hope that psychological states are causally efficacious in virtue of their psychological properties is tantamount to the hope that psychological states are causally efficacious in virtue of their intentional properties.

There is, however, a *prima facie* problem in understanding how the intentional properties of mental states could be causally relevant to the production of other mental states and behavior.[5] The problem arises as a result of two metaphysical commitments that together appear to preclude intentional properties from being causally efficacious. One metaphysical commitment is that intentional properties are not physical properties. This commitment has received wide support as a result of certain arguments against the unity of science thesis of, e.g., Oppenheim and Putnam (1956); in the context of psychology, this unity of science hypothesis was expressed as the claim that mental properties are type-identical to physical properties. The arguments against the

type-identity of physical and psychological states purport to show that mental states (and other kinds of states) are functional in nature, and are therefore multiply realizable; this multiple realizability then shows that these higher level properties cannot be connected to lower level properties by exceptionless bridge laws (i.e., higher levels cannot be 'reduced' to lower levels, cf. Nagel, 1966). Therefore, the higher level mental properties cannot be identical with the lower level physical properties (see, e.g., Putnam, 1975, and Fodor, 1975). The second metaphysical commitment is that the physical universe is causally closed; that is, to the extent that physical events are caused, they are caused only by other physical events.[6]

It is easy to see how these two commitments are incompatible with the claim that the intentional properties of mental states are causally efficacious. If the only kinds of properties that participate in causal relations (i.e., if the only properties that are causally relevant) are physical properties, and intentional properties are not physical properties, then pretty clearly intentional properties do not participate in causal relations. Thus, there is at least a *prima facie* mystery about how mental states could cause anything (e.g., other mental states and behavior) in virtue of their intentional properties.

One approach to removing this mystery appeals to the notion of *supervenience*. Heil (1992, ch. 3), following the work of Kim (e.g., 1990) and others, has argued cogently that the supervenience relation is a real part of the natural order. The supervenience relation seems to occupy a middle ground between the identity and distinctness of two properties. It is normally expressed, in this context, as the claim that there is no difference in mental (or intentional) properties without a difference in physical properties. Mental properties are not wholly distinct from physical properties in that they arise or emerge from combinations of physical properties and can only be altered by altering the physical properties that they arise out of (and hence supervene on); but nor are intentional identical to physical properties insofar as many distinct physical properties can provide a supervenience base for the same mental property (and given the transitivity of identity, if these physical states were identical with the supervenient mental property, and that mental property were identical with a different subvenient physical property, then, contrary to fact, the two different physical properties would be identical).[7]

Supervenience is thought to contribute to a resolution of the mental causation mystery in the following way. If mental events can be shown to supervene locally on physical events of one kind or another, and those physical events are firmly rooted in the causal order, then the sense in which mental events are causally relevant will be no more obscure than the sense in which other supervenient events or properties

(e.g., economic trends or political climates) are causally relevant. In all such cases, the supervenient properties are not distinct from the physical properties, so the supervenient properties partake in the causal relevance of the physical properties on which they supervene.

Robb (1997), among others, has pointed out, however, that the supervenience solution to the problem of mental causation does not quite work. While mental properties are not distinct from physical properties, they are still not identical to physical properties. Thus, if only physical properties are causally relevant, then any properties that are not identical to physical properties are not causally relevant. Moreover, just what is it for a property to be 'not distinct but not identical to' another property? Why should we believe that there is such a bizarre metaphysical category? And, relatedly, the supervenience relation does not provide us with a mechanism that explains how properties of this supervenient sort manage to be causally relevant. Without some general understanding of supervenience and supervenient causation, it does no good to appeal to other cases of apparent supervenience and supervenient causation in defending the causal relevance of the mental; perhaps all such cases involve no causal relevance at all.

There is, however, a refinement of the supervenience solution to the problem of mental causation that fills in some of the gaps left by the supervenience account as it has been put forward. The refinement is found in Robb (1997, p. 186ff),[8] and it begins by pointing out a pernicious ambiguity in the original formulation of the problem. The term 'property' admits of at least two distinct interpretations. On the one hand, a property is a universal (or class, for the nominalist); it is what all instances that have a given property share (cf. Plato's One-Over-Many argument). On the other hand, a property is an abstract particular, or trope (see Williams, 1966, and Campbell, 1990):

> "On this reading, properties are particulars, wholly present in the individuals that instantiate them but logically incapable of being (at the same time) wholly present elsewhere." (Robb, 1997, p. 186)

Two objects may have the same color, and in that sense they share a property. But in another sense, one object's tokening of that property is distinct from the other object's tokening of that property. Thus, in one sense they share a color property, but in another sense they do not. Following Robb, I will refer to the former sense by the word 'type', and the latter sense by the word 'trope' (and where it is necessary to conflate the two, I will use the term 'property').

Claims about tropes can illuminate questions about individuation of the sort at issue in this book. I have often remarked that individuation is about stating the properties that a token must have if it is to be a token of a given kind. In talking about the color of an object, for example, we can see this question as the question of what (physical) properties an object must have if it is to have a particular color. One way to understand this question is as a question about what (physical) trope (or tropes) an object must have if it is to have the color trope (or tropes) that it has. That is, the question of individuation can be understood to be a question about the particular features a particular object or entity must have if it is to have the feature being individuated. In the case of the color of an object, the object must have certain physical tropes that reflect light in some particular way. In this case, the physical trope in virtue of which the object reflects light as it does is *identical* to the color trope the object possesses. Unlike supervenience relations, which are cast in terms of types, discussions of tropes concern concrete particular objects and the particular features, tropes, they possess. Thus, the appeal to tropes allows us to express the sense in which supervenient types are distinct (they belong to distinct classes or fall under distinct universals), and the respect in which they are not (the token members of a supervenient class share particular tropes with the token members of the subvenient class).

This refinement is crucial in understanding the original formulation of the problem of mental causation. In the original statement of the problem, we appealed to the metaphysical commitment that mental (specifically, intentional) properties are not physical properties. Recall that the chief reasons for sustaining this commitment derive from the multiple realizability arguments against the type-identity of mental and physical properties. But, in the language that we are now using, these arguments support only the claim that mental types are not identical to physical types; that is, multiple realizability shows *only* that mental states *qua* mental, and physical states *qua* physical, belong to different classes, or fall under different universals. These arguments do *not* show that mental tropes are not identical to physical tropes. As with colors, the physical and color types might stand in a supervenience relation (such that color types may be multiply realizable by physical types), and yet it may be that particular physical and color tropes are identical. In the case of mental types and physical types, they may stand in a supervenience relation, and yet physical tropes might be identical with mental (or intentional) tropes. Thus, in light of the distinction between types and tropes, we must reformulate our first metaphysical commitment to be the claim that mental types are not physical types. This is, however, perfectly

consistent with claim that mental tropes are identical to physical tropes.

The second metaphysical commitment can also be reformulated. It is generally assumed that causal relations hold in virtue of the properties of the causal relata. This, of course, is ambiguous between type and trope construals. Construed as a claim about types, this is just the claim that causal relations relate members of different classes in nomologically consistent ways. Construed as a claim about tropes, it is the claim that particular objects or events stand in causal relations in virtue of the tropes that each possesses. This construal captures what Horgan (1991, p. 88) refers to as "the here-and-now" character of causation; it is in virtue of the fact that this particular object has at this time and place this particular trope that on this occasion it causes that particular object to have at that time and place that particular trope. Thus, when we talk about the causal closure of the physical, we can understand that to be the claim that only physical tropes may have causal relevance to the production of effects.

Given the two ways of formulating each of the metaphysical commitments that give rise to the problem of mental causation, we are in a position to see how those commitments are perfectly consistent with the claim that mental (re: intentional) properties are causally efficacious in (or causally relevant to) the production of other mental states and behavioral output. While it is true that mental types are not physical types, it is perfectly conceivable that mental tropes are identical to physical tropes. The commitment to causal closure, on the other hand, is the commitment to the claim that only physical tropes can cause behavioral effects (which are assumed to be characterizable by their physical tropes as well). But if mental tropes just *are* physical tropes, then there is no problem in understanding how the mental is causally relevant to the physical: The mental just *is* the physical, when we are talking about tropes, even though mental types are not identical to physical types.[9]

The only mystery left here is how it could be that physical tropes are identical to mental tropes.[10] But this is, really, little more than the age-old question of how there could be mental (or intentional) properties in a physical world. Any account of the mental, and *a fortiori*, any account of mental causation, must accept the existence of the mental even if we are left to speculate (albeit in a constrained and principled way) about the ontological relation between the mental and the physical. Thus, to the extent that we can have an account of mental causation, casting that account in terms of the supervenience of mental types on physical types, and ultimately the identity of mental and physical tropes, appears to hold much promise.

2.4 Mental Causation and Anti-individualism

Now that we have some grasp on how mental causation might occur, I will show, through use of the repaired version of the Fodorian argument, that anti-individualism is inconsistent with our best hope of a solution to the mental causation puzzle. While Robb (1997, p. 179) is agnostic about whether the trope account of mental causation conflicts with anti-individualism, I believe it can be shown that there is indeed a conflict.

Our account of mental causation substantiates step (1) in the repaired Fodorian argument. Step (2) is supported by the general metaphysical point extracted from the previous chapter, and laid out in section 2.2 above. Step (3) is just a statement of anti-individualism. I will now show that a defense of step (4) can be extracted from the trope account of mental causation. Thus, my argument in this section may be construed as a defense of CPN that rests not upon STCP, but upon the only plausible account of mental causation we can hope for. Given the support for the premises, the conclusion will follow. Ultimately, I hope this section will articulate the problem that anti-individualism faces concerning mental causation that many before me (in particular Fodor, but also Block and others) have tried but failed to articulate.

Let us assume that causally efficacious physical tropes are not relational; that is, let us assume that the physical tropes in virtue of which an object or entity enters into causal relationships are contained wholly within the objects that constitute the causal relata. This should not be a contentious assumption in the context of discussions about anti-individualism if only because Burge himself (1989, 1993) is willing to allow that causally efficacious physical properties are intrinsic and not relational. If those tropes are to be identical with mental tropes, as they would have to be for the mental (or intentional) to be causally efficacious, then, I claim, the mental tropes cannot be relational, either. Here is why: If mental tropes could be relational in the way that anti-individualism suggests, then it would be possible to have the same physical trope being identical with two distinct mental tropes, provided only that the relation that is individuative of the mental trope in question is changed appropriately. Imagine, for example, a slow switching case of the sort discussed in chapter 2. In such a case, the physical trope that is token identical with the content, or intentional trope 'water is wet' later becomes token identical with the intentional trope 'twater is wet'. Given a plausible principle of the transitivity of token identity, we would get the consequence that the intentional trope 'water is wet' is token identical to the intentional trope 'twater is wet'. For the intentional trope 'water is wet' is identical to physical trope P, and P is identical to the intentional trope 'twater is wet', so 'water is wet' is identical to 'twater is wet'. But this, of course, is contrary to

what anti-individualism maintains. Thus, if content is to be causally efficacious, according to the trope account, then it cannot be relationally individuated in the way that anti-individualism claims that mental content is individuated relationally.[11]

One might object, at this point, that it is not just the mental tropes that change in slow switching cases, but also the physical tropes. The internal trope P on Earth, the objection runs, is replaced by an exactly similar internal trope P' on Twin Earth; the difference between P and P' is that P is identical with the relational trope consisting of P in the context of water, while P' is identical with the relational trope consisting of P' in the context of twater. Since P is not in any sense token identical to P', there is no way to transfer the identity of 'water is wet' to the content 'twater is wet' through the trope P. Thus, the unwelcome consequence for anti-individualism never arises.

There are several difficulties with an objection of this sort. First, it is simply implausible to maintain that the internal tropes have changed in the trip from Earth to Twin Earth. Indeed, apart from some wildly mysterious view about tropes popping in and out of existence, the only sense in which the internal tropes might be said not to be token identical is if their identity is temporally indexed; in this case, the trope at t_1 on Earth is not identical to the trope at t_2 on Twin Earth simply because t_1 has passed, and it is now t_2. This is an implausible way to individuate tropes in this context; in discussing causal efficacy, it is taken for granted that a trope is causally efficacious at a particular time, but it is allowed that numerically the same trope could be causally efficacious at a later time. This is only possible if tropes can be individuated independently of their time of occurrence. But even if one individuates tropes relative to their time of occurrence, there is nothing in this practice to prevent the tropes from *also* being individuated independently of their time of occurrence. According to this sort of individuation, the internal physical trope P on Earth is token identical to the internal physical trope P on Twin Earth. Thus the problem for anti-individualism arises all the same. That tropes might be divided up in ways that prevent the problem from arises does nothing to deflect divisions of tropes in ways that allow the problem to arise. To provide a real objection to the problem facing anti-individualism, one must show not that tropes can be divided up in ways that prevent the problem from arising, but that tropes *cannot* be divided up in ways that allow the problem to arise.

Another difficulty with this objection is that it assumes that tropes can be internal and relational at the same time. In effect, the objection claims that an internal physical trope is also a relational physical trope in virtue of being embedded in a particular context. This is implausible, and appears to rest on a serious misunderstanding of the

trope proposal. It is perfectly coherent to talk of a token internal to a system as belonging to both an internal type and a relational type; my brain states may be individuated in terms of internal properties or in terms of relational properties. But this entails no similar plausibility in maintaining that an internal trope can also be a relational trope. Tropes are abstract particulars possessed by tokens; they are, as it were, the instantiations of types. The tropes that characterize the internal nature of an entity are simply *different* from the tropes that characterize the entity's relations to its environments. A token entity may be type individuated internally and relationally, and similarly it may possess both internal tropes and relational tropes; but it makes no sense whatsoever to talk about a single trope being type individuated internally and relationally. When we are talking about tropes, being internal and being relational are mutually exclusive.[12]

I do not believe there is any way to resist the conclusion that anti-individualism is inconsistent with our account of mental causation. It is worth pointing out, however, that Heil (1992, pp. 138-146) deploys tropes in defense of an anti-individualist account of the causal relevance of mental properties. Heil can be read as making the following sort of criticism of my discussion so far. In talking about the causal relevance of the mental, I have helped myself to a notion of causation. But this notion stands in need of analysis. At least part of that analysis, Heil says (p. 140), can be achieved by appealing to "...a system or network of counterfactual and subjunctive conditional truths..." that help specify the types of dependency that defines causal relationships. In the case of S's mental states and her behavior, the types of counterfactual conditionals that are applicable include those that specify what would happen if things were slightly different (i.e., what would happen "...in a range of nearby worlds..."). So, for example, S's intention to drink water caused S to drink water only if it is the case that if S did not form the intention to drink water, she would not drink water.

Now, I might interject at this point the claim that, given anti-individualism about contents, we can imagine twin scenarios on which S would perform the same behavior even if she lacked the intention to drink water; for example, suppose S spent some time on Twin Earth but has now returned and forms instead the intention to drink twater. It would seem that the counterfactual stated above is false of S. If the truth of that counterfactual is at least a necessary condition on S's content playing a causal role in the production of her behavior, then, given an anti-individualist individuation of content, it seems we have reason to doubt the causal relevance of S's intentional content.

Heil escapes this problem by pointing out that twin scenarios involve worlds that are physically quite remote from the actual world; in order for S to lack an intentional content involving water, we had to

consider a world involving Twin Earth and travel between Earth and Twin Earth, etc. Such worlds are far different from the actual world. But the truth of counterfactuals, as Lewis has shown us, is assessed relative to *nearby* possible worlds, not remote possible worlds. Thus, it is irrelevant to the truth of the counterfactual in question what would happen in one of these remote possible worlds; thus, Twin Earth considerations are perfectly consistent with the causal relevance of S's intentional content, and are therefore irrelevant to the question of whether the content of S's intention is causally relevant. Indeed, according to Heil (1992, p. 147), any physical changes (including changes to S's history) that affect the content of S's intention would also affect S's causal powers. This is because there is no nearby world in which S would have the same physical constitution but lack the same intentional content.

Heil concludes that the trope account of mental causation is consistent with anti-individualism; thus, anti-individualism entails no more of a mystery about mental causation than individualism. But even assuming that Heil is right that anti-individualism is not defeated by an appeal to a counterfactual analysis of causal relevance, the earlier problem remains: In order for the intentional to be causally efficacious, there must be a sense in which it is identical to the physical (even if there remains a sense in which the two are not identical). But if we assume that it is only the internal properties of the physical that are causally efficacious (an assumption that is tantamount to the claim that causation is local, and is hence an assumption that anti-individualists do not deny), then it follows that the intentional must be internal, too, on pain of the contradiction that two distinct contents would be identical. Thus, Heil's attempt to show that anti-individualism is consistent with a trope account of mental causation fails to address the problem I have raised in this section. Thus, in spite of Heil's discussion, we have every reason to believe that anti-individualism is committed to an implausible view of mental causation.

The deeper problem for anti-individualism is quite general, and not limited to the particular trope account of mental causation defended here. If mental properties are to be a part of the causal order, and the causal order, insofar as it is constituted by local causal relations, respects only internal properties as causally relevant, then mental properties must also be internal to the states that have them. Since anti-individualism does not respect this constraint (indeed since it is defined by the *refusal* to respect this constraint), it cannot provide a plausible account of the causal relevance of the mental. Thus, I claim, there is at least a *prima facie* difficulty in reconciling anti-individualism with the causal relevance of the mental.

Another way to put the point I am urging is that the anti-individualist is committed to the false claim that there are causal powers that are relational in just the sense that the anti-individualist individuation of content is relational. In defending the assertion that causal powers are not relational, I have defended CPN, and hence step (4) in the repaired version of the Fodorian argument. Burge, however, has tried to undermine step (4), and defend the claim that there are relational causal powers. He claims that there is a tradition of sciences that recognize the legitimacy of relational causal powers; so, psychology would face no more of a problem of causal relevance than any other science that also countenances relational causal powers. Indirectly, then, this challenges the purported difficulty for anti-individualism that I have developed in this section; if it really were a difficulty, it would apply to large numbers of special sciences, which is manifestly implausible. In the next section, I will address Burge's considerations along these lines.

3. *The Argument from Scientific Taxonomy*

In this section, I will lay out and then criticize central aspects of Burge's case for the claim that there are relational causal powers. Burge grants STCP in order to show that CPN is false. There are, he claims, many examples of sciences that appear to taxonomize states and entities by appeal to their relational properties. If these taxonomies are by causal powers, as STCP states, then it appears that CPN is false; the causal powers appealed to in at least some sciences are indeed wide. But if CPN is false, then there is no support for (4) in the Fodorian argument. Thus, an anti-individualistic psychology committed to causal powers that are wide in the same way that wide content is wide would not be alone in taxonomizing states relationally; indeed, its relational taxonomy would receive the sanction of standard taxonomic practice in science. However, as I shall argue, insofar as Burge's argument rests on STCP, it is mistaken. Moreover, and perhaps more importantly, even if Burge were right that sciences taxonomize by appeal to causal powers, the respects in which the kinds in those taxonomies are relational are crucially different from the respects in which causal powers in an anti-individualistic psychology would have to be relational. Thus, I will in this section present two different ways to resist Burge's claim to have justified the postulation of relational causal powers; there is nothing in the appeal to scientific taxonomy that will save anti-individualism from the case presented against it in section 2.

3.1 *Wide Scientific Taxonomy*

It is central to Burge's thinking on this issue that we let our actual explanatory practice guide our metaphysical predilections (cf.

Burge, 1993, esp. p. 118). These sentiments are echoed by Egan (1991, p. 190) and Wilson (1992, esp. p. 133), though it is not so clear that they are prepared to parlay these considerations into an argument for wide causal powers in the way that Burge does.[13]

Burge cites several different examples of taxonomies that appear to be relational. Land masses are taxonomized in geology by their relations to other land masses. If those relations had been different (if, e.g., they did not slide around on the face of the earth), they would not be taxonomized as land masses. And, significantly,

> "...the causal powers associated with these land masses would have been differently described by geology. But the chemistry and physics of the relevant land masses need not (metaphysically need not) have been affected. A land mass with substantially the same non-relational physical features could -- because of its different relations to the environment -- be of a different geological kind. Thus, geological kinds do not supervene on the kinds of masses that are described by physics and that constitute geological entities." (Burge, 1989a, p. 309.)

Much the same lesson, Burge claims, can be drawn from biology. Organs, such as the heart, are taxonomized by appeal to their role within the system in which they participate. A heart just *is* the sort of thing that pumps blood. A microstructurally identical organ that pumps waste would not be a heart. And, significantly,

> "The causal powers attributed to such an organ by biology would be different from those attributed to a heart. ...The biological kind *heart* does not supervene on the chemical structures of the material that constitutes hearts." (Burge, 1989a, p. 310.)

If there is no problem with the relational taxonomies of these sciences, Burge claims, then there should be no problem with a relational taxonomy in psychology of the sort proposed by anti-individualism. After all, he says,

> "psychology's relation to neurophysiology is similar, in the relevant respects, to the just mentioned relations between natural sciences." (Burge, 1989a, p. 310.)

So, if there are wide causal powers corresponding to the wide taxonomies of geology and biology, there should be no difficulty in accepting wide causal powers corresponding to the wide taxonomy of anti-individualism in psychology.

Wilson's (1992, pp. 116-121) examples from anthropology, sociology, and evolutionary biology may be taken to make the same point. In anthropology, the taxonomy of taboos is relational (cf. p. 116). In sociology, the taxonomy of criminals is relational (cf. p. 117). In evolutionary biology, the taxonomy of species is relational (cf. p. 118), as is the taxonomy of genes (cf. McClamrock, 1995, p. 30). Egan (1991, p. 189) points to the same sorts of examples. On the assumption that STCP is true (an assumption Egan and Wilson are reluctant to make), such relational properties are causal powers. Insofar as an anti-individualistic psychology participates in this tradition, its relationally taxonomized states may be legitimate causal powers.

McClamrock (1995, esp. chs. 1-3) articulates a framework for understanding these sorts of taxonomies that allows us to see how causal powers might be relational. As Fodor pointed out long ago (e.g., Fodor, 1975, Introduction; see also Lycan, 1987, chs. 4 and 5), many varieties of system can be partitioned into different levels of causal organization. Kinds at higher levels are often *multiply realized* by kinds at lower levels; that is, lower level differences sometimes make no difference to the taxonomy of a system at higher levels. For example, different electronic configurations can implement the same computational transition. McClamrock also notes that kinds at lower levels are often *context-dependent*; that is, the same lower level kind sometimes participates in the realization of different higher level kinds. For example, if coding mechanisms are allowed to vary, the same DNA sequence can realize different genetic properties.[14]

Adapting the familiar probabilistic notion of *screening off*, McClamrock (following Brandon, 1984) shows how one might go about determining the "appropriate" level at which to explain a given phenomenon. An event A screens off another event B from an effect E iff the probability of E given A and B is equal to the probability of E given just A. This screen off test (SOT) can be represented like this:

$$\text{(SOT)} \qquad P(E/A\&B) = P(E/A)$$

The application to the present concern with levels of causal organization is relatively straightforward. If variations in lower level properties (within the class of properties that realize the higher level property) has no impact on the probability that the effect will occur, then the appropriate level of explanation is not the lower level, but is instead the higher level. When explaining a particular computer printout, for example, many different low-level machine languages can implement the same higher level command to print; the low-level variations are therefore screened off from the effect, so long as they implement the appropriate higher level command. On the other hand,

when variations in higher level properties are irrelevant to the probability that the effect will occur, while low-level variations make all the difference, it is then the lower level at which the appropriate explanation for that effect is to be found. When explaining why a falling computer killed a bug, it is irrelevant what program the computer was running at the time. All that matters to the explanation of the effect is the set of lower level physical properties that, quite incidentally, implement that program.

In terms of this framework, we can understand how causal powers in the aforementioned taxonomies might be relational. When explaining the presence of phenotypes, genetic properties screen off lower level properties. Given our interest in genetic properties, we must type-identify DNA sequences from the point of view of genetics. From this perspective, however, the same DNA sequence will lead to different phenotypes in the context of different coding mechanisms or different locations on the genome. Thus, as far as genetics is concerned, the causal powers of DNA sequences, what phenotypes the DNA are able to cause, are dependent on their context of occurrence; vary the coding mechanisms or the location of a sequence on the genome and there will be a corresponding variation in the resulting phenotype.

Similar stories can be told for the other sciences mentioned above. Causal powers can be individuated relationally when viewed from the perspective of a particular higher level taxonomy whose kinds are realized by context-dependent kinds at a lower level. From the perspective of psychology, neural events have relational causal powers insofar as their causal powers depend on their history of relations to states of affairs in the environment (see McClamrock, 1995, e.g., p. 30).

3.2 A Crucial Disanalogy

Let us, for a moment, imagine that anti-individualists are right to think that the sciences mentioned in the previous section taxonomize states relationally. Even given this point, Burge's attempt to place an anti-individualistic psychology within a tradition of sciences that appeal to relational causal powers, and McClamrock's attempt to explain how these causal powers could be relational, fail to show what they set out to show. The relations involved in the taxonomies mentioned above are of a relevantly different character from the relation of interest to anti-individualists.

In the case of biology, for example, Burge argues that hearts are organs that are type-identified relative to a certain context; hearts are organs that pump blood and occupy a certain functional role within an organism. The causal powers of hearts are bound to this context. But notice that the context relative to which a heart is type-identified is contemporaneous; in other words, it is not a history of relations to a

particular context that bestows upon hearts the causal powers that they have. The causal powers of hearts, and hence the type-identity of a heart, depends on the relation of a particular organ to its *current* context of operation. The same holds true for the cases from genetics and geology.

But the sort of context relative to which anti-individualism type-identifies contents, and hence the context of interest to the anti-individualist, is *past* context. We have the contents we do, according to Burge, in virtue of having acquired our concepts in a certain way in a certain environment. He says, for instance, in considering epistemological problems associated with wide (non-individualistic) content, that one's

> "...thoughts would not switch as one is switched from one actual situation to another twin actual situation. The thoughts would switch only if one remained long enough in the other situation to establish environmental relations necessary for new thoughts." (Burge, 1988b, p. 652)

He also articulates just how one's social community might be relevant to the individuation of the meanings of terms in one's idiolect (i.e., one's concepts); for example:

> "The individuation of our concepts and meanings is sometimes dependent on the activity of others from whom we learn our words and on whom we depend for access to the referents of our words." (Burge, 1989b, p. 187)

According to Burge, then, it is a fact about one's history, that one has learned one's concepts in a particular social environment, that is relevant to the individuation of one's concepts and the attitude contents in which they figure. If these contents are to be causal powers (in any sense, but in particular in the sense described in section 2 above), then the causal powers have to be type-identified by appeal to those same historical properties. But none of the examples from science that anti-individualists provide involve type-identification relative to historical properties. So, even if certain sciences do identify relational causal powers in a particular domain at a particular level of organization, they do not identify relational causal powers of the sort that would help the anti-individualist. As far as anti-individualists have shown, then, there is no tradition of appropriately similar relational taxonomies in science in which an anti-individualistic psychology would participate. There is no other science that type-

identifies causal powers in the historical way that the anti-individualist must.[15]

We do not have to look far to see what is significant about this disanalogy. The notion of screening off, which McClamrock uses to identify the appropriate level of causal explanation, was originally used by Salmon (1984) as a way of understanding certain aspects of causality and causal relevance at a single level of organization.[16] When we apply the screen off test (SOT) to, say, McClamrock's case of explaining a phenotype by appeal to DNA and coding mechanisms, we find that the context (i.e., the coding mechanisms) cannot be screened off from the phenotype by the DNA; both affect the probability of the effect, so both are causally relevant factors.

But the situation is different when we turn to the anti-individualists' historically taxonomized mental states. Mental states, according to the anti-individualist, consist in neural events with a particular history of relations to the environment. In stark contrast to the case in genetics, however, the neural properties of the event appear to screen off the historical properties from the effect.[17] The probability of a given behavior given a neural event with a history of relations to the environment is equal to the probability of that behavior given the neural event without such a history. The fact that the neural event has some history of relations to the environment is irrelevant to its causing of the behavior; it would cause that behavior no matter what its relational history has been, no matter whether it has had such a history at all.[18]

This does not entail that these causal histories are irrelevant to the production of behavior. For it is plausible to suppose that we would not have the thoughts we have without such histories. Therefore, we would not exhibit the behaviors we do without those histories. But this sort of causal relevance suggests nothing about relational causal powers; the same sort of causal relevance is found in physics, where any causal powers are manifestly not relational. The cue ball, for instance, would not have struck the 8-ball unless it was itself struck by the cue. This does not entail that the causal powers of the cue ball are context dependent in any interesting way. The cue ball, like a thought, will cause what it does no matter what its history might have been, so long as it has the current properties that it has (including, of course, vector-properties). Thus, so long as taxonomy in science is by causal powers, histories will not be relevant to taxonomy.

Burge does claim (1989, p. 30) that causal histories can be relevant to taxonomy in science, and hence to individuation in metaphysics, in another way. Igneous rocks, for example, are taxonomized in geology by appeal to their origins, not what they can cause; this does not mean that taxonomy in geology is non-causal, it

simply means that it is causal history that is relevant to taxonomy, not causal consequences. Perhaps the same sort of taxonomy is suitable for psychology.

If this were indeed all that Burge were claiming, there would be little to dispute. But Burge clearly does mean to make a stronger claim in arguing that there are causal powers appropriate to the wide content of anti-individualism. This is revealed in his repeated insistence that psychology should be able to determine for itself what are to be legitimate causal powers in its universe of discourse (see, e.g., Burge, 1989a, p. 306).[19] Thus, the fact that causal histories can be appropriate sources of taxonomy in science does not insulate the argument from scientific taxonomy from the criticisms I have levied against it. Even if we allow that some causal powers are relational, SOT helps us to see that this will be of no help to the anti-individualist. Relational causal powers of the sort that would provide a supervenience base for wide content, or a suitable physical trope with which the wide content is token identical, receive no sanction from the relational character of taxonomy in science, even if we accept STCP and the structure of the argument from scientific taxonomy.[20]

3.3 A Further Implication of the Disanalogy

We have so far seen that the causal commitments of anti-individualism are not commensurate with standard appeals to causality in science. Owens (1993) has argued that these causal commitments make it impossible for ordinary propositional attitudes to be individuated in terms of their content, and play a causal role in the production of behavior. Saidel (1994) has argued that the tension Owens purports to discover is illusory. He suggests that causal powers, no less than content, can be individuated widely; and, moreover, the reasons that compel one to accept that contents are wide are precisely those that would compel one to think that causal powers are wide as well. Once this is recognized, he claims, there will no longer be any reason to suppose that propositional attitudes cannot figure in causal explanations of behavior. This is an interesting discussion, one to which the disanalogy at issue in the previous section has crucial application. I will in this section deploy the aforementioned disanalogy to save Owens' argument from Saidel's criticisms.

The argument under discussion is this:

> 1'. Folk psychological explanation makes essential explanatory appeal to propositional attitudes, to states individuated in terms of content (the representational thesis).

2'. Folk psychological explanation is in the business of offering causal explanations of behavior (the causal thesis).

3'. Explanatory accounts which are in the business of providing causal explanations for the behavior of various systems should distinguish between the explanatory states of such systems only if they differ in causal powers.

4'. Hence, folk psychological explanation should distinguish between the psychological explanatory states of the twins, Alf and Alf* (their beliefs, and so on) only if they differ in causal powers.

5'. Since the twins are physically identical there is no difference in causal powers between their corresponding explanatory states.

6'. Hence, folk psychology should not individuate in such a way as to distinguish between the (ordinary) psychological explanatory states of the twins, that is, it should not individuate in such a way as to justify attributing a particular explanatory state to one twin and not to the other.

7'. Hence, belief and desire should not be individuated in such a way as to justify attributing differing beliefs and desires to such twins (by 1' and 6'). (Owens 1993, p. 248)

The argument is interesting because (7') is inconsistent with anti-individualism, according to which twins are said to have beliefs and desires (and propositional attitudes) with different contents; hence, by (1'), the twins are to be attributed differing beliefs and desires.

Though there is much to be said about the argument, Saidel concentrates his attention on (1') and (5'), which, he claims, are ambiguous. In (1'), the use of "content" is ambiguous between wide and narrow construals. Saidel claims (pp. 659-60) that those who favor wide content (presumably for reasons deriving from Twin Earth thought experiments) will accept (1'), but only on the understanding that it concerns wide content, while those who favor narrow content (and hence presumably object to Twin Earth arguments) will accept (1'), but only on the understanding that it concerns narrow content.

This is debatable, however. One may be committed antecedently to the view that the propositional attitudes of folk psychological explanations are individuated by appeal to their contents, and then discover, by considering Twin Earth arguments, that the contents are wide. Indeed, this seems exactly to be the way of reasoning of Burge and those whom his arguments have convinced. It is

not that they antecedently determined that contents are wide, and then determined that these are the states one should appeal to in folk psychological explanations. Thus, accepting (1') is largely independent of whether one thinks that contents are wide or narrow.

It remains true, however, that one could reinterpret (1') to involve one's favored view of content, so, strictly speaking, (1') could be seen to be ambiguous. But it is not obvious that this observation advances the dialectic; it is clear that Owens is supposing that arguments derived from Twin Earth thought experiments are compelling, so it is clear that he intends for (1') to support the view that contents are wide. If he is wrong about that, then of course there is no tension between wide content and narrow causal powers; there will be no wide contents for narrow causal powers to conflict with.

The real interest in Saidel's discussion, however, comes not with the claim that (1') is ambiguous, but with the claim that (5') is ambiguous. In (5'), Saidel claims, it is the use of "causal powers" that is ambiguous between wide and narrow construals. Anti-individualists are compelled to think that contents are wide by considering how intentional properties of a given internal state vary with contexts; they should likewise be compelled to think that causal powers are wide, Saidel claims, by considering the effects of a given internal state across different contexts (see Saidel, 1994, p. 661).[21] Saidel supports this contention by providing an example where the same object has different effects in different contexts (his vertical jump on Earth is two feet, while on Jupiter it would be six inches). Thus, if one accepts that contents are wide, one must accept that causal powers are wide as well; and if one accepts that contents are narrow, one must accept narrow causal powers as well. Either way, the tension that Owens sees between individuation by content (which he presumes is wide) and individuation by causal powers (which he presumes are narrow) never arises.

Owens, of course, disagrees, but to do so, he must provide some justification for holding that in cases where contents are wide, causal powers can be narrow. Saidel argues that no such argument is forthcoming, but I think his arguments miss their mark. I think that Saidel is mistaken on at least two counts. First, as far as Saidel has shown, there are no legitimate wide causal powers, so, as far as he has shown, endorsing wide content does not mandate endorsement of wide causal powers. Second, and following on the point made in the previous section, Saidel's argument that endorsing wide content can mandate endorsement of wide causal powers equivocates on the notion of context relative to which width of content and causal powers is to be assessed.

Take first the notion of wide causal powers. Since there is no question that some causal powers are narrow, the real question is whether there are, in addition, some causal powers that are wide. If

narrow causal powers are limited to the internal properties of individuals, then wide causal powers are those that are not limited in this way. The claim that causal powers are internal entails that there is no difference in causal powers without a difference in internal features. Thus, to be committed to the claim that causal powers are narrow is to be committed to the claim that two twins, A and B, will have the same causal powers. To be committed to the claim that causal powers can be wide, then, is to be committed to the claim that A and B, despite being internal twins, can have different causal powers. So, to determine whether causal powers are narrow or wide, we need to determine whether the causal powers of A and B are the same or different.

Fodor (1991) argues that the way to test whether A and B have different causal powers is to see whether they have different effects across all nomologically possible contexts.[22] It is not a legitimate test of whether A and B have different causal powers, for example, to consider what effects A's thoughts have on Earth and what effects B's thoughts have on Twin Earth. Even the individualist agrees that, e.g., the effect of A's thoughts is that she will drink water while the effect of B's thoughts is that she will drink twater; individualism is the view that their *thoughts* are the same, not that their environments contain the same types of liquids. A's thoughts and B's thoughts *would* have different causal powers, however, if they would result in the drinking of different fluids even where A and B are in the same context. Obviously, however, this will not happen. It is, in part, the obviousness of this conclusion that explains why Owens is willing simply to *assume* that causal powers are not wide.

The example Saidel considers, however, is not relevant to whether causal powers are wide, because it is not relevant to whether two internally identical objects have the same effects across contexts. The example instead involves different effects in quite distinct contexts, which is consistent with causal powers being narrow. If A has effect E1 in context C1 and E2 in C2, and B has E1 in C1 and E2 in C2, then A and B have the same causal powers as each other, even while each has a different effect in one context from what he has in another context. Saidel's example does not show anything about whether causal powers supervene on intrinsic features of the object in question, so it does not show anything about wide causal powers.

Though this is only a preliminary point, it does put pressure on Saidel's argument. For suppose it is possible for content to be wide. Saidel's argument purports to show that an endorsement of wide content compels an endorsement of wide causal powers. But if there are no wide causal powers, Saidel's argument would compel anti-individualists about content to embrace a fallacious view. In this case, there would

indeed be significant tension inherent in the relation between wide content and causal powers; adherents to wide content would be required to accept a falsehood.

My most important disagreement with Saidel's argument, however, concerns the claim that endorsing wide content compels endorsement of wide causal powers, while endorsing narrow content compels endorsement of narrow causal powers. Saidel's argument for this claim depends on the assumption that, as far as anti-individualism is concerned, the contexts that are relevant to the individuation of content are those that are relevant to the individuation of causal powers. I believe that Saidel shares this assumption with Burge, but as we saw in the previous section, both are mistaken. The contexts relevant to the anti-individualistic individuation of content are historical, while the contexts relevant to causal powers are not.

Saidel, however, never identifies clearly the precise way in which Burge's anti-individualism involves one's historical and environmental contexts. He says only that

> "[t]he Twin-Earth thought experiments argue that beliefs supervene on internal physical states plus some external factor (either history, or local environment, usually)."[23] (Saidel 1994, p. 663-4)

The parenthetical disjunction reveals an imprecision in Saidel's understanding of anti-individualism. Attention to such infelicities might seem to be of little general interest. But as a matter of fact, this point is key. For when we consider the sort of context that is relevant to the individuation of causal powers, we see that the distinction between historical and current contexts becomes crucial.

Saidel, like Owens and Burge before him, considers the example of the heart. Burge (1989b, p. 312) assumes, let us suppose correctly, that hearts are individuated in terms of the functional role homologous devices have played in the evolutionary history of hearted species. Saidel claims that it follows from this that the causal powers of a heart *qua* heart are tied to that history, so the causal powers of a heart include only pumping blood, since that is the effect hearts have had throughout their evolutionary history. It may be, of course, that the causal powers of a particular token heart, *qua* physical object, include more than pumping blood. For *qua* physical object, hearts can do much more than pump blood; they can, for example, pump waste or squash bugs if dropped from the top of a ladder (cf. Burge, 1993, p. 101). But that just shows that we have to be careful how we type identify the object whose causal powers we are interested in assessing. To consider the causal

powers of a heart *qua* physical object is not to consider the causal powers of a heart *qua* heart.

> "If we are interested in the causal powers of the heart, and what makes it a heart is its history, then we cannot escape that history in our enumeration of its causal powers. If, however, we are interested in the causal powers of the object, then we can escape that history." (Saidel, 1994, p. 663)

But, Saidel claims, it is the causal powers of the heart *qua* heart that interest us.

Saidel then applies this lesson to the case of psychology. The belief that p, for example, is composed of an internal state that bears a relation to some contextual factor or another. If we take the internal state out of its context, we will be removing that relational property in virtue of which, according to anti-individualism, it is the belief that it is. Thus, when we try to compare the effects of the belief across contexts, we fail to do so since it is not the belief *qua* belief whose effects are compared across contexts; it is only the internal state (or belief *qua* internal state) whose effects are considered. As far as anti-individualism is concerned, "...to ask about the causal powers of a belief is to ask about the causal powers of a physical state with a particular history." (Saidel, 1994, p. 664) When we do that, we discover that

> "[w]e cannot make identity claims about the causal powers of a belief in different contexts precisely because we do not think it is the same belief in different contexts." (Saidel, 1994, p. 664)

Thus, the sort of context test that Fodor prescribes for determining whether or not the causal powers of two internally identical states are the same, and hence whether causal powers supervene on intrinsic states of an individual or object, begs the question against the anti-individualist by assuming that it is even possible (even counterfactually in thought experiments) to put the same belief in different contexts. And without the context test, the individualist has no grounds by means of which to deny the existence of wide causal powers.

This is an ingenious move, to be sure, but I do not believe that it succeeds. First, notice that Saidel's claim that we are interested in the causal powers of beliefs and hearts in their constitutive contexts is questionable, and insures that his argument will fail to motivate wide causal powers. The question about the width of causal powers is a question about whether or not the causal contribution of twin beliefs and hearts to the production of their effects is limited to their internal states alone, or to the internal states in conjunction with the relevant

contexts. To determine this, we must determine whether or not the *internal* states have the same effects across nomologically possible contexts, not whether the same internal/context pairs have the same effects across those contexts. Compare this to the Twin Earth thought experiments involving contents; there the internal states are held constant, while the context is varied. *That* is the way to show whether or not something, be it content or causal powers, is limited to internal states.

The latter test, involving internal/context pairs, *is* impossible, as Saidel points out, but it is also irrelevant to the very point at issue. The individualist grants that, according to anti-individualism, different contexts beget different contents, but questions whether these differences are reflected in the causal powers of the states. To answer this question, we must determine whether causal powers are limited to internal states alone, or internal states plus contexts. For if they are limited to internal states alone, then the differences in the contents of internal states wrought by the variations in context will not be matched by a corresponding difference in causal powers; that is the problem to which Owens gives expression. As it turns out, of course, internally identical twins *do* have the same effects across all nomologically possible contexts, thus demonstrating that causal powers are limited to the internal states of individuals even if contents are not, and thus giving rise to Owens' tension.

Moreover, even if Saidel were right that the determination of wide causal powers would require the assessment of whole (internal/context) belief states across contexts, and that this is impossible, he would gain no argument for wide causal powers. Whether or not there are wide causal powers depends on whether or not the causal powers of a thing include internal states plus contexts, and not internal states alone. According to Saidel, however, it is impossible to determine whether causal powers vary with contexts because contexts are not allowed to vary. We have at most, then, the absence of a test for causal powers, not a test that shows wide causal powers. Thus, we remain without reason to suppose that there are any wide causal powers at all.

But even were we to concede these points to Saidel, a familiar and deeper problem threatens his argument. While it is true, according to anti-individualism, that some changes in context affect the identity of a belief content (and hence a belief), we saw above that, by Burge's own admission, those contexts are very sharply circumscribed. It is *only* a variation in the history of one's relations to the (physical or linguistic) environment that will affect the individuation of the contents in question. A variation in history alone, however, will not call for a variation in causal powers.

Consider a thought experiment, similar to the quick switching case Burge (1988b) considers, where Alf is transported in his sleep to Twin Earth. Suppose that both he and Alf* think the thought that each would express with the word forms "water would be good right now", and both subsequently utter, as a result, "Bring me some water"; each then receives some twater. Imagine also the switch going the other way, with corresponding results. Have we not here just applied the context test, and learned that the causal powers of Alf's and Alf*'s thoughts are the same, and hence that those causal powers supervene on the internal states of the individuals? I suppose that Saidel might protest that we have not considered the effects of Alf's water thought because, in placing Alf in the Twin Earth context, we have changed his thought from a water thought to a twater thought. But such a protest would be misguided, as Burge himself would readily admit (indeed Burge, 1988b, p. 652, insists on just this point!). Alf's internal state has retained its history of relations to Earthly environments, so it continues to constitute Alf's water thought. It is not, after all, impossible to consider Alf's and Alf*'s beliefs across the relevant contexts (e.g., Earth and Twin Earth), thus it is not impossible to determine whether the causal powers of Alf's and Alf*'s thoughts are different; thus it is not impossible to determine whether the causal powers of Alf's and Alf*'s thoughts are limited to factors internal to Alf and Alf*.

The reason for this error in Saidel's argument should by now be apparent. The contexts on which the anti-individualist supposes contents depend are quite different from the contexts relative to which causal powers are assessed. Contents, as we saw above, depend on *historical* relations to environments. But causal powers are assessed relative to *current* relations to environments. It is *of course* true that historically individuated states cannot be considered in different historical contexts, and still retain their identity; but they can be considered in different current contexts, and yet retain their identity. Indeed, as we have seen, it is possible for two internally identical states, individuated relative to different historical contexts, to be considered in the same current context. When we do this, it is no surprise to find that their effects are the same. In fact (as we will see throughout this chapter), it is easy to see that no variation in historical context ever bears on an internal states' causal powers; historical relations are simply not relevant to the assessment of causal powers. The current context of an individual does, of course, affect what the individual can cause. But that is precisely why the effects of twin states must be assessed across all nomologically possible current contexts; such a procedure allows us to distinguish the causal contribution of the individual from the causal contribution of the individual's current context.[24] The fact that we do discover that the two contributions are

separable shows that the causal powers of the individual are limited to the internal states of the individual.

Thus, for those keeping track, the score is this: Both current contexts and internal states are relevant to the production of effects, but by varying current contexts and internal states in thought experiments, we can disentangle their separate causal contributions; historical contexts, on the other hand, are never relevant to the production of effects. That, in fact, is what gives rise to the tension between the historical individuation of anti-individualism and the ahistorical individuation by causal powers.

We must be careful to notice, however, that since the contexts that purportedly determine wide contents are historical, and not current, we need not alter the (anti-individualistic) type identity of an individual's thoughts when we consider them across current contexts; again, different histories call for different contents, on anti-individualism, but not for different causal powers. Thus, even by Saidel's own lights, his argument fails to resolve the tension between the individuation of propositional attitudes by wide content, and the individuation of the attitudes by causal powers.

Saidel's reply to Owens fails because he is mistaken in thinking that adherence to wide content has any bearing on whether one should adhere to wide causal powers; *mutatis mutandis* for narrow content and narrow causal powers. Saidel errs in this matter, I suspect, largely because he is incautious in his appeals to the contexts relative to which the width of contents and causal powers are assessed. Indeed, careful attention to such contexts suggests that the notion of wide causal powers is a fiction. Owens' original tension is therefore likely to remain threatening to those who maintain that anti-individualism will fit comfortably in a folk psychology that is in the business of providing causal explanations of intentional behavior.

3.4 Against Relational Causal Powers

The sense in which an anti-individualistic psychology would taxonomize relationally is, I have argued, relevantly different from the way in which other sciences taxonomize relationally. But even if this disanalogy did not exist, there is still a serious flaw in the way that the argument from scientific taxonomy is framed. As I suggested above and in the previous chapter, there is no reason to expect that regularities and generalizations in science are framed in terms of causal powers. As far as empirical methods of discovery and verification are concerned, regularly co-occurring properties will not be distinguished even when, metaphysically speaking, there is a difference between them. Thus, the causal joints of nature cannot simply be read off the surface of scientific taxonomies.

Indeed, given the local nature of token causation, it is natural to read the apparently relational taxonomies of science to which Burge and other anti-individualists point such that no evidence of relational causal powers is forthcoming. If all token causal instances are local in nature (i.e., non-relational), then all types of those token instances should be similarly local. For if we assume that any causation between physical events is local, my argument in section 2.4 demonstrates that causation at higher levels *must* also be local (on pain of being committed to the claim that no higher level properties are causally efficacious). If certain special sciences taxonomize relationally, it must then be because they are not taxonomizing by appeal to causal powers alone. Instead, we may suppose that such taxonomization exploits certain regular co-occurrences of properties. In particular, such taxonomies can be seen to exploit the rather pedestrian fact that events of the sort that interest the special sciences tend to occur in the same sorts of contexts. Psychological events, for instance, tend to occur within a stable physical and social environment (though see chapter 2, section 2.2, for arguments to the conclusion that these environments are not as stable relative to individuals as anti-individualists like to think). This does not mean that the causal powers of psychological events are relational; rather, it suggests only that the causal powers of psychological events tend to be exercised within relatively stable and unchanging contexts. The distinction between the event and its context remains even if the event and its context tend to co-occur. So, since there is no particular reason to think that taxonomy in science appeals to causal powers alone, Burge cannot rest his resistance to CPN (and step (4) in our argument) on STCP; for STCP is, in general, false.

We can see, then, that even if Burge was (apparently contrary to fact) intending to argue for causal powers that are relational in the sense suggested by McClamrock's analysis, his point would not go through. For McClamrock's notion amounts to the familiar claim that causes require background contexts if they are to result in effects. This is not a claim that anyone would be wise to doubt. Indeed, McClamrock has provided us with a framework that allows us to talk with a certain degree of rigor about distinctions between causes and the background conditions they rely on in the production of effects.

Probabilistic relevance of the sort appealed to in SOT is at least a minimal condition on causal relevance (see, e.g., Eells, 1982, pp. 223-4). Thus, SOT helps us to see that from the perspective of genetics, there are at least two kinds of lower level causal factors involved in the production of phenotypes. Both DNA and coding mechanisms are causally relevant to the production of phenotypes (surely other factors are relevant as well). DNA cannot cause a phenotype without coding mechanisms, and coding mechanisms cannot cause a phenotype without

DNA. But there is more to our notion of causal powers than can be captured probabilistically. Causation is about making things happen, and it is customary to focus on the dynamic precursors of event as its cause, rather than the static background contexts within which the dynamic precursors operate. Pragmatic factors, such as explanatory interest and prior knowledge, can sometimes compel us to view the more stable background context as the cause.[25] In the normal case, given some prior knowledge of coding mechanisms, we try to isolate the DNA sequence responsible for the production of a phenotype. Had we instead some knowledge of DNA sequences, we might view coding mechanisms as the salient cause against a background context that includes certain DNA sequences.

Notice, then, that there is no particular reason to claim that DNA has causal powers that are relational, unless one is willing to claim that *all* causal powers are relational. It is true, of course, that there are certain things that DNA can cause only in the presence of certain background contexts that include some specification of coding mechanisms. But billiard balls require certain background contexts to effect the particular motions they do in other billiard balls. To claim that the causal powers of a billiard ball are relational is either to trivialize the notion of relational causal powers, or to reduce it to absurdity. The fact that the effects of DNA are modulated in some way by the prevailing background context does not make the causal powers of DNA any more relational than the causal powers of billiard balls, whose effects are also modulated by other causal factors.

It is no doubt true that certain causal factors regularly co-occur, so that there are counterfactual-supporting generalizations involving taxonomizations that lump together two or more causal factors. At bottom, this is what Burge is pointing to in making claims about relational causal powers. Though entities structurally identical to hearts can contribute to the production of many different effects in many different contexts, the fact that these sorts of structures occur only in conjunction with a distinctive embedding context (i.e., circulatory systems of living organisms) warrants restricting our attention to its causal powers relative to that normal context. As Burge himself puts it,

> "...in asking for the causal powers of the heart, we implicitly expect physiological patterns of properties to be cited. We do not expect citation of powers that would be studied by physics. Pumping blood is usually considered relevant; squashing a bug if dropped from a ladder is not." (Burge, 1993, p. 101.)

But by restricting our attention to the effects of a thing in its normal environment, we are, either explicitly or implicitly, neglecting

the full range of its causal potential. Burge (1989a, p. 310) is quite right to point out that "the special sciences track causal patterns that cut across each other." But there is no obvious inference from this fact to the conclusion that the special sciences deal in causal powers that are relational in a way that, say, the causal powers of physics are not. Higher level sciences simply *hold fixed* certain causal factors while others are allowed to vary. They begin with, e.g., structures type-identified at a lower level of organization, and attend to the causal powers of such structures only in some normal context. This is not the recognition of yet another causal power that was not recognized when the entity was viewed from the perspective of the lower level science; and it is certainly not the recognition of an additional relational causal power that was not visible from the point of view of the lower level science. Higher level sciences simply restrict the range of causal powers considered in a way that, for example, physics does not. Insofar as the special sciences track causal patterns that cut across each other, they merely partition the space of causal powers of lower level structures, and do so in different ways.[26] Biologists do not care about the causal powers of certain configurations of muscle tissue in free fall, but only about the causal powers that muscle tissue has when it is embedded in its normal organismic context.

When the causal powers that result from lumping two or more regularly co-occurring causal properties together are related nonlinearly to the component causal powers (that is, when combinations of causal powers give rise to effects that are more than the mere sum of what each is capable of in different contexts), we should be prepared to countenance a new, higher level of causal organization. The chemical structure of DNA bestows upon it causal powers that are in certain ways more than the mere sum of the causal powers of its constituent chemicals. But higher order causal powers are not thereby relational causal powers. There is the causal relevance of one chemical constituent, the causal relevance of another, and a higher order causal power that results from the co-occurrence of the two chemicals. Similarly, there is the causal relevance of the DNA, and the causal relevance of the coding mechanisms, and possibly the higher order causal power that is some function of the causal powers of DNA and coding mechanisms. But, from a theoretical standpoint, there is little to be gained, and much to be lost, by claiming that the causal powers of DNA are context-dependent in any substantial way (i.e., in a way that ordinary physical causal powers are not). Indeed, there is *less* relativity to context in the higher level sciences insofar as the range of contexts relative to which causal powers are assessed is more circumscribed than it is in the lower level sciences.[27]

So, given that STCP is itself false, and that we can now explain precisely why it is false, I conclude that the argument from scientific taxonomy provides no support for relational causal powers. Moreover, not only is it illegitimate to appeal to the surface of scientific taxonomies to enumerate causal powers, even if we could make such an appeal, there is a serious disanalogy between the way in which other sciences taxonomize relationally and the way an anti-individualistic psychology would.

4. The Argument From Behavior

One way Fodor proposes to defend CPN is by appealing to the context test (CT):

> (CT) Token events C1 and C2 have the same causal powers iff their effects are the same across all nomologically possible contexts. (Cf. Fodor, 1987, p. 44.)

The point of CT is that the causal powers of two tokens must be compared in such a way that any differences in effect cannot be attributed to context.[28] To get a sense of the import of CT, imagine the familiar twin story in which S and Twin S reside in environments that are distinguished only by the microstructure of the stuff that is in the lakes and streams; as usual, the twins would be unable to distinguish the Earthly substance, water, from the twin substance, twater. In such environments, S's use of the word forms "Bring water" causes someone to bring water, while Twin S's use of those word forms causes someone to bring twater. This is not a legitimate comparison of the causal powers of S and Twin S since the effects considered do not cross all the relevant contexts. Instead, the effects of S's utterance are considered in the context of S's environment, and the effects of Twin S's utterance are considered in the context of Twin S's environment. A test such as this would recommend a difference in causal powers even where none exists; for even S's effects in the two environments would differ in spite of the fact that her mental states are the same. If we place S and Twin S in the *same* environment, however, say S's, we can see that the effects of their utterances would *not* be relevantly different; both utterances would result in the arrival of water.[29] According to CT, then, the causal powers of S and Twin S, in particular the causal powers of their mental states, are the same. So, individuals cannot differ in their causal powers without also differing internally; and since (or if) neuroscience taxonomizes brain states by appeal to such internal states, it follows that individuals cannot differ in their causal powers if they do not differ in their brain states. And because there is no difference in causal powers, the causal powers are revealed to be narrow; the environment is

discovered to contain *additional* causal factors that shape the effects of a particular deployment of S's causal powers. So, given the support of CPN, premise (4) is secure and the conclusion is forthcoming.

In this section, I will consider several attempts to rebut Fodor's use of CT. I will conclude that they are uniformly ineffective, and that CT stands as a plausible philosophical tool in the assessment of causal powers.

4.1 Misunderstanding the Twins

Burge's initial reaction to CT is to chastise Fodor for applying CT only in contexts where the remarks of just one twin will be misunderstood (cf. Burge, 1989a, p. 311). He therefore imagines a case where S and Twin S produce utterances, and both are *understood* by hearers who comply with their requests (for water and twater, respectively). Thus, in the same context, the twins' mental states will have different effects. Here, he claims, the twins will have been shown to have different causal powers by Fodor's own CT.

In point of fact, however, the twins will not have been *shown* to have different causal powers; it has merely been *assumed* that they have insofar as it has been assumed that they are in different psychological states, and that this difference in psychological states is (in the cases Burge would prefer to consider) sufficient to produce different states of mind in the hearers. If the point of CT is to determine if the twins are indeed in different psychological states (by determining if their causal powers are the same or different), then it cannot be assumed, in the very description of the case, that they are in different states.

Of course, Burge could be imagining a set-up where S speaks to a group of hearers who understand his utterances, because they are Earthlings, and Twin S speaks to a group of hearers who understand his utterances, because they are Twin Earthlings. These contexts are the same insofar as they are both contexts in which the speakers' utterances are understood. But this is inadequate. CT demands comparisons of effects across *all* nomologically possible contexts where the twins are placed in the *same* context. It is not sufficient that they be in contexts that are the same in some one respect, but different in others; they must be in contexts that are the same in *every* respect. The case we are supposing that Burge imagines is not like that since the context in which S is understood is in many respects different from the context in which Twin S is understood. This sort of case does not satisfy the condition set forth in CT, so it is irrelevant to CT and could hardly count against it.

4.2 Intentional Individuation of Behavior

Suppose we allow intentional identification of mental states in the ways suggested by Burge's wide content arguments. Consider this challenge to CT: Microstructurally identical mental states might be, if Burge's wide content arguments are successful, intentionally distinct; and behaviors are often identified relative to the intentional states that cause them; e.g., S's trip to the refrigerator might be water-seeking behavior, while Twin S's would be twater-seeking behavior. Since CT proposes to identify causal powers by appeal to effects, and because the effects (behaviors) in these two cases would be distinct, CT would be forced to return different verdicts regarding the Twins' causal powers.

This is, essentially, the challenge that exercises Fodor in his (1991). Fodor responds to it by denying the legitimacy of relational causal powers that involve conceptual links between cause and effect (cf. 1991, p. 24). Wilson (1992, pp. 127-33) criticizes this move. Though I am not persuaded by Wilson's criticisms, I will not digress to consider them here (I will consider a similar though more developed challenge in the next section). Rather, I think we have available to us an alternative response. Fodor is driven into his move in part by the fact that he grants the scientific legitimacy of many relational causal powers. I do not. We have seen no clear examples of them in science, and there is no further reason to accept them. Thus, were Burge to insist on postulating relational causal powers as a result of his wide content arguments, he would be forced to admit that an anti-individualistic psychology is anomalous with respect to all other sciences, something he is explicitly unwilling to do. By rejecting wide causal powers wholesale, we force Burge either to accept that an anti-individualistic psychology would be a scientific anomaly, or to renounce this sort of deployment of the wide content arguments in the context of a discussion of causal powers.

Moreover, the purported intentional difference between the twins' intentionally characterized behavior is something that CT can correct for. Even if we grant for the sake of argument that the contents of S's and Twin S's contents are distinct, it does not follow that there is no intentional description of their behaviors on which those behaviors are identical. When S and Twin S are placed in the same Earthly environment, they both drink water. Moreover, there is a description of this activity under which S and Twin S both drink water intentionally: Both S and Twin S think of the water in the glass before them in the same way (e.g., as the clear and tasteless liquid that one can find in the lakes and streams), and both have the intention to drink that liquid. Thus, both S and Twin S do what they do intentionally. Thus, we do not have to resort to descriptions of behavior in terms of physical movements to find a description under which S and Twin S behave identically. Thus, CT is not forced to concede that the twins behave

differently in the same context, and hence that their mental states have different causal powers.

The anti-individualist may claim that because Twin S conceives of her behavior differently (as twater drinking behavior), her behavior should be given a different intentional description. But Twin S's conception of her behavior would be wrong, so it would wrong to describe her behavior differently from S's, even intentionally; both engage in water drinking behavior. Burge would be right to say that, according to anti-individualism, B is *trying* to drink twater. Therefore, according to anti-individualism, B's beliefs and desires cause her to try to do something different from what S's beliefs and desires cause her to try to do. But this again simply *assumes* an anti-individualistic individuation of mental states, and hence does not *show* that the causal powers of the twins' mental states are different. Regardless of what the anti-individualist says Twin S is trying to do, she does the same thing as S when placed in the same context. Therefore, the causal powers of their mental states are the same. If the anti-individualist claims that the contents of those states are not the same, then those contents are not the causal powers of the mental states (and those contents are hence not, as it were, trope identical to the causal powers of the states in question), contrary to what the anti-individualist is trying to argue.

Similar remarks apply if S comes from a linguistic community that, say, uses the term "drink" differently. The anti-individualist may again want to say that she conceives of her behavior differently, and therefore that the behavior should be individuated differently; Twin S is tw-drinking while S is drinking. But, again, in the context provided, Twin S's behavior would not differ from S's; both would again be drinking water (as that behavior would be described in English). One could still insist on describing the behaviors differently on account of the different ways in which (according to anti-individualism) the twins conceive of their behaviors, but such different descriptions could be given no matter how each conceives of her behavior, and even if they conceive of their behavior in the same way. Such an insistence entails nothing about the individuation of behavior. So far as we have seen, there is a perfectly ordinary intentional description of the twins' behaviors on which they behave identically. So, by application of CT, we have no grounds to distinguish between the causal powers of the mental states of S and Twin S.

4.3 The Adams Family

The claim that S and Twin S really do behave differently, even in the same context, has been pressed by Adams (1991, 1993) and his colleagues (e.g., 1992) in a series of papers defending so-called broad content.[30] These arguments also appeal to the purported intentional

distinctions between the behavior of the twins, but they also involve many other points; for this reason, I would like to consider them rather carefully to see if they present any difficulty for the proposed use to which CT is being put.

The Adams Family begins with the claim that psychology is in the business of explaining intentional behavior. They then assert that, when described intentionally, the behaviors of the twins is distinct. Indeed, if one assumes that anti-individualism is right about contents, then there will be descriptions of the behaviors on which they differ; i.e., S intentionally drinks water while Twin S intentionally drinks twater (or unintentionally drinks water, if she is also on Earth). Thus, it is simply not true that the twins have the same effects across all nomological contexts. Even if Twin S is on Earth, she is not intentionally drinking water; S, on the other hand, is intentionally drinking water. So, then, should not psychology provide different explanations for these two different types of behavior?

Surely not, I claim. Just what is this sense in which the behaviors of twins are distinct? The Adams Family repeatedly refers to the behaviors in question, the sorts of behaviors, they claim, that it is the business of psychology to explain, as "intentionally drinking water" and "intentionally drinking twater", etc. If we put these behavioral descriptions in the form of infinitives, however, we can begin to see that there is something pretty obviously fishy about this way of describing the explananda of psychology. Let us take the description of S's behavior: 'To intentionally drink water' (the infinitive form of 'S intentionally drinks water'). Grammatically speaking, of course, this is a split infinitive, and there are good reasons why those interested in using the language properly avoid such mistakes. In this case, the reason is that the split infinitive description makes it look like the so-described 'intentionally drinking water' is a different action from drinking water. Stated properly, however, the temptation to treat them as different actions is noticeably lessened: Drinking water intentionally *is* drinking water; it is the drinking of water that is brought about (caused) by the intention to drink water. When water is drunk intentionally, it is not a *different kind* of water-drinking from a water-drinking that is not done intentionally. It is, rather, the *same* water-drinking in both cases. What is different, of course, is the *cause* of the water-drinking; in the one case it is the intention to drink water, and in the other case it is something else (perhaps some other intention).

Now, it may be that psychology would want to classify behaviors by their causes, in which case, water-drinking that is done intentionally would be different from water-drinking that is not done intentionally.[31] But this, of course, presupposes that there is a classification of the behaviors that is independent of their causes; in

the case under discussion, for example, the twins both engage in water-drinking behavior. Whether this behavior is done intentionally or not is, of course, and question of interest to psychology, but the very question requires that there be a way of characterizing the behaviors independently of their causes. Thus, it is simply not true that psychology's interest in explaining so-called intentional behavior requires a classification of behaviors by their intentional causes. It may perfectly well (and indeed must) classify behaviors independently of their intentional causes, even while it concerns itself only with those independently characterized behaviors that are caused by the intentions to do them.

So, when the Adams Family claims that psychology is in the business of explaining why S 'intentionally drinks water', they are mistaken in assuming that this is some sort of special behavior distinct from the water-drinking that someone might do unintentionally. Had they instead said just that psychology is in the business of explaining why S drinks water when that behavior is done intentionally, there would be little temptation to think of drinking water intentionally as a distinct *behavior* from drinking water unintentionally; it is the same behavior in each case, but done as a result of different causes. Characterized in this way, the goal of psychology is at once clearer and more general. Some psychologists might be interested in explaining S's drinking of water by appealing to the intention to drink water, while others might be interested in explaining how S came to have the intention to drink water. And among those who are interested in explaining how S came to have the intention to drink water, there are some who might look for a rationalizing explanation (e.g., S felt thirsty and believed that drinking water would quench her thirst), and some who might look for a functional explanation (e.g., S came to have the intention as a result of the interaction among several desires, the interaction among several beliefs, and the interaction of the output of the desire mechanism with that of the belief mechanism).

The only respect left in which the twins' behavior is different, then, is the respect in which, according to anti-individualism, S's behavior is the result of the intention to drink water (i.e., it is the drinking of water that is done intentionally), while Twin S's behavior is the result of the intention to drink twater (i.e., it is the drinking of water that is not done with the intention to drink water). The individualist is perfectly happy to allow that this is how the anti-individualist would like to describe the case, and that this description embodies a coherent way of classifying the behaviors. The problem for the anti-individualist, however, is that it concedes that the behaviors that, according to anti-individualism, are caused in different ways are nonetheless the *same* behaviors. Thus, the anti-individualist is

committed to accepting that S and Twin S behave the same way; they simply disagree on whether the behaviors were caused in the same way. In fact, the Adams Family appears to concede this point:

> "Bring Twin-Jerry to Earth, and he will exhibit the same behavior as Jerry. They both will exhibit H2O-drinking behavior. But they will do this because of different learning histories and different belief-contents." (Adams, 1991, p. 149.)

But whether the behaviors were caused by the same intention (or belief) is, of course, the very question at issue. And the way I am proposing to settle that question is by determining whether the states have the same causal powers. The way we are determining whether they have the same causal powers is by appealing to CT. Now, we have seen that there is no warrant for the assertion that S and Twin S engage in different behaviors simply because one does what she does intentionally while the other (according to anti-individualism) does not. The behaviors are the same, as far as anyone has shown. So, according to CT, S's intention and Twin S's intention have the same causal powers. But if they have the same causal powers, then given that metaphysical individuation is by causal powers, two states with the same causal powers will be individuated in the same way. So, if S's and Twin S's intentions have the same causal powers, they are the same kind of state (from the point of view of metaphysical individuation). And given the account of mental causation defended in section 2, the anti-individualist can resist this conclusion only on pain of not having an account of mental causation.

The Adams Family might try to argue for Dretske's (1988) process (or 'component') view of behavior on which behavior is the process of an internal state causing, say, some bodily movements (see, e.g., Adams, 1991, 1993, for an endorsement of this view). On this view, the intention is part of the behavior, in which case a difference in intention entails a difference of behavior. Given this view of behavior, it is trivially true that, for the anti-individualist, twins will have different behaviors even when placed in the same context. The Adams Family suggests that we should accept this view because it will allow us to explain how broad content can be causally efficacious.

Though a complete discussion of this process view of behavior cannot be undertaken here, there are a couple of points that should be raised against its deployment in this dispute. First, if broad content is to be construed consistently with individualism, as I maintain, then there is no motivation for the process view; we already have an account of how broad content that is trope identical with internal states of the individual can be causally efficacious. There is no need for the process

account to save broad content. Second, if broad content is to be construed along anti-individualistic lines, then we fall back into the problem that there is no reason to think that such historically relational properties are ever causally efficacious. Third, rather than cling to the unintuitive process view of behavior, why not simply hold that psychology is interested in more than behavior? It is also interested in how people come to be set up so that states that are about, e.g., water, tend to cause water-drinking behavior. Despite frequent (perhaps behavioristically inspired) claims to the contrary, there is no reason to think that psychology is exhausted by its interest in explaining behavior.

There is a fourth problem that is a bit more intricate. The process view of behavior is supposed to allow for the causal efficacy of content in the following way. If behavior is an internal state's causing certain movements, then the content of that internal state explains (presumably causally) why it is involved in the production of that movement; it is because this state represents water, for example, that it is involved in the production of water-drinking movements. But the state need not have this broad content to explain why it has been recruited to cause water-drinking movements. It may instead represent the state of some disjunction of retinal states that is itself reliably correlated with the presence of water. So, the purported broad content of the state is, strictly speaking, not necessarily for the explanation of behavior construed as a process; narrow content will suffice.

This is a problem for the Adams Family for two reasons: First, the process view of behavior, they claim, is required if broad content is to be explanatory. But even given the process view, there is no need for broad content. So, the Adams Family has not made good on its promise to show that narrow content is inadequate to the explanatory purposes of psychology. Second, the manner in which broad content comes to arrange a system so that the internal state that has that content causes some appropriate bodily movement is *also* supposed to explain how the state comes to have its broad content; it is because the state is correlated in some way (and hence indicates, in Dretske's language) some external state that it is recruited to cause the appropriate movements. But if this is the story, then it would appear that the state would already have to have its broad content if it is to be recruited to cause the bodily movement. It is because my intention is about water that it is recruited to cause water-drinking behavior. But if this is the case, then Dretske's entire theory of content acquisition, which is what he invokes the process view to help justify, falls short of its goal. Either the content is already there as that in virtue of which the state is recruited (in which case Dretske's theory assumes what it is supposed to show), or it is not (in which case having that content is not explanatorily necessary for explaining why the state is recruited to cause the movement).[32]

It should be noted, however, that while the Adams Family is sympathetic to the process view of behavior, they do not appear to be sold on it. In Adams (1991, p. 139), for example, it is claimed that the process view is inessential to an account of the explanatory role of broad content. Thus, we should not expect the Adams Family to press too hard for the process view as a way of resisting our defense of CT.

In fairness to the Adams Family, they view the issue in an importantly different light. They have bought into the assumption that individualism is true only if the contents of mental states are not about any matters outside of the individual. They, following many others, call this narrow content. Broad content, then, is content that is about matters external to the individual. The Adams Family is mainly concerned to show that narrow content is neither necessary nor sufficient for the purposes of psychological explanation.

In chapters 1 and 4, however, I distinguished between narrow content, on the one hand, and content that supervenes on, or is trope identical with, internal properties of the individual, on the other; similarly, there is a distinction between broad content and content that does not supervene on, or is not trope identical with, internal properties of the individual. One can maintain, I argued, that mental content supervenes on properties internal to the individual, but nevertheless makes reference to or is about matters outside the individual. The fact that a representational state is about matters outside of the individual is not what would make anti-individualism true. Rather, anti-individualism is true only if states are about matters outside of the individual *in virtue of* the fact that the states are related to the external world in one of a number of well-specified ways. If it should turn out that mental states have their intentional properties in virtue of local features of biological brains, then, even though they are about external matters (and in that sense broad), individualism would be true.

Thus, I have no bone to pick with the Adams Family over the necessity of narrow content; the truth of individualism, the claim I want to investigate, is independent of the issue narrow content. But I do want to make one point about broad content and CT. Given their interest in showing that broad content is all the content that is needed in psychological explanation, the Adams Family claims that two different broad contents may be equivalent with respect to causal powers. S's and Twin S's causal powers are different because they developed their concepts in different contexts. If you bring them to the same context, their causal powers might turn out to be equivalent. They claim that there is nothing about this possibility that is inconsistent with using broad content to explain behavior within its original context (see Adams, 1993, pp. 44-5). Presumably the Adams Family believes this because they also believe that psychology is only concerned with

individuals in their normal contexts (such that psychology can taxonomize states differently in different contexts despite the equivalence of their causal powers across contexts).

I disagree with this view for a variety of reasons. First, why think the general supporting principle about what psychology is concerned with is true? It appears simply to be assumed since it is not even made explicit, and no effort is made to defend it. This would not be particularly damaging so long as there is no reason to doubt it. But, given our discussion in chapter 2 of the prevalence of slow switching, there is good reason to doubt that it is true. As people pass in and out of different linguistic communities and physical environments over the courseof their lives, they pass in and out of the very contexts that the Adams Family wants to relativize psychology to. But there is no reason whatsoever to think that psychology should be relativized in this way. If the causal powers of two states of an individual are the same across such contexts, then, it would seem, psychology should reflect this similarity on pain of drawing distinctions where there are no differences. As the Adams Family remarks, this would be no problem for broad contents of the anti-individualist sort if such context crossing were rare, but given that it is not rare, this does pose a problem for this anti-individualistic view of the relationship between content and causal powers.

Second, given that our issue is a metaphysical one, and given that metaphysical individuation is by causal powers even if psychological taxonomy is not, equivalence of causal powers is all that we require for equivalence of metaphysical individuation. Thus, while the Adams Family may be right to argue that broad content is sufficient for psychological explanation even if distinct broad contents can have the same causal powers, this cuts no ice on the question of metaphysical individuation. Thus, whether the Adams Family is concerned with the metaphysical issue or not, the fact remains that if anti-individualism is committed to distinguishing metaphysically between two states with equal causal powers (and it is), then anti-individualism is committed to a false metaphysical individuation of mental states.

I conclude, then, that no claims from the Adams Family about the individuation of behavior cause trouble for the use of CT that I am defending. Perhaps because of quite different interests, the arguments of the Adams Family do not affect my argument for CT.

4.4 Context Device

Burge claims that the reliance on understanding the twins and the intentional individuation of behavior is unnecessary. I will now consider another objection to CT that does not involve assumptions of the sort we have so far been concerned with. Burge asks us to imagine

> "...a device that traced the histories of individuals, recording whether they had been in causal contact with [water]. Such a device could bring [water] to an individual with such a causal history when he made the sounds "Bring [water]" -- and not otherwise. In such a context, [S] would have different effects from [Twin S]. Once again there is a possible context in which the twins' acts produce different effects." (Burge, 1989a, p. 311.)

Burge entertains a possible objection to this example (i.e., that the operative causes in the example are not relevant to psychological explanation), but it is not the objection I want to press.

Instead I would like to argue that Burge has misidentified the relevant causal tokens being tested by CT in the example provided. Burge assumes, without argument, that the relevant tokens are those mental states of S and Twin S that result in their utterances. But this is false. It is not just S's and Twin S's current *utterances* that interact with the device in question; it is their current utterances *and their histories.* That is, the causally relevant tokens, in Burge's example, are *temporally extended*; each twin has causally interacted with the device over an extended period of time. The consequences of the twins' utterances are the cumulative effect of each twin's behavior over time. Most importantly, these temporally extended tokens are *not internally identical.* Since S has interacted with water and Twin S has not, the device is clearly responding to microstructurally different, temporally extended tokens. Thus, this is not an example of a twin-style thought experiment at all; it is, therefore, an inappropriate application of CT and could not possibly count against it.

It is crucial to a proper understanding of this response to Burge that we realize that individualism is not committed to the simple minded view that multiple causal interactions over time are somehow illegitimate. Nor, for that matter, are effects always to be understood as unextended snapshots. Threshold phenomena belie such a naive view of causal powers, and vector quantities belie the correspondingly naive view of effects. The individualist *is* committed to holding that sequences of states or events that are microstructurally or internally identical have identical effects, and hence identical causal powers. So, if each twin has the same history of interactions with his environment *and with the device that measures that history,* then the individualist is prepared to maintain that they will have the same effects in all contexts, and hence the same causal powers.

It is also crucial to Burge's example, and my response to it, that the device have causal commerce with the twins during the time preceding their utterances. If the device had only measured the final

momentary utterances of the twins, it would not be able to tell any difference (as the microstructures of the utterances are identical). Thus, the effects would be the same, as would, by CT, the causal powers.

This response also suggests a further response to Burge's initial objection to CT, which involved putative understanding on the part of the hearers. I do not grant that the contents of the twins' mental states are distinct; but if I did, I would also maintain that for such understanding to occur, the hearers *must* have had access to the twins' different histories of interaction with water and twater; there is no other way they could understand the twins' utterances differently (thus, the case in which they understand the difference without exposure to the twins' histories is not a nomologically possible context). In the case of the hearers, then, like the case of the device, the relevant tokens whose causal powers are being assessed are temporally extended. And in such cases, there are no differences in effects in the face of microstructurally identical causes; but since that is the only kind of case that would conflict with CT and make trouble for individualism, there is nothing here for the individualist to object to (microstructurally distinct causes result in distinct effects).

4.5 Historical Individuation

Might Burge simply agree here that psychological states are historically individuated states? Was it not his point all along that the individuation of a psychological state depended on a history of relations to a particular type of substance or linguistic community? These questions are misleading; there remain crucial distinctions between his vision of psychology and that of the individualist. First, I doubt very much whether Burge wants to maintain that token psychological states are temporally extended in the sense operative in my response to his objection to CT. Rather, he wants to maintain that *occurrent* tokens are individuated by psychology in part by appeal to their histories of relations to their environments. In the case I am describing, however, the relevant token state is not just the occurrent state, but a temporal extent that begins with the first interaction of the agent with the hearer (or the device), and continues up till the most recent interaction. Thus, the relevant token does not merely have the property of having *had* a particular history, it *is* that history. Since the internal nature of this history is different for the different twins, and since this is what the hearer (or device) has interacted with, we no longer have a case to which CT applies; CT applies only to tokens whose internal natures are identical.

Moreover, the point in the individualist response to the device example was never to suggest that psychological states are temporally extended. As Burge notes himself, the case has little to do with

psychological explanation of the sort at issue. My point is that even if it did, it stands as no objection to CT.

4.6 CT and Scientific Taxonomy

Burge (1989a, p. 312) also suggests CT is insensitive to differences among the typologies of different sciences, and should be rejected for this reason. The waste-pumping organ that is homologous to a heart, for example, would be taxonomized differently from a heart by physiologists despite the fact that the two would have the same effects when placed in all the same contexts. But this is what we should expect, given the interpretation of scientific taxonomies developed in section 2 of this chapter (and throughout the previous chapter). Different sciences attend to the causal powers of homologous entities *within* different normal contexts. CT, on the other hand, is about assessing the effects of internally identical states or objects *across* all nomologically possible contexts. The different taxonomies do not *disagree* about the causal powers of the entity, they are simply concerned with different subsets of those causal powers. Thus, the fact that CT ignores the different ways that sciences taxonomize is, in my view, *to its credit.* Lower level sciences tell us what causal powers there are, and the different special sciences merely choose from among them.

But there is another respect in which CT accords perfectly well with the taxonomic practices of the special sciences. The causal powers of homologous entities will be typed identically by a given science, according to CT. If physiology is interested in the waste-pumping causal powers of an organ in the context of waste, it will be interested in the waste-pumping causal powers of an homologous organ in the same context. The fact that another science may be interested in certain other causal powers of such an organ is no embarrassment at all for CT; it is no surprise that the same state or object can have different effects in different contexts. This is, again, an unsurprising attitude to take if one also believes that the space of causal powers is broader at the lower levels, and that the special sciences are merely in the business of partitioning that space in various systematic ways.

5. Concluding Remarks

It is worth considering what individualists find objectionable about the anti-individualistic view of causal powers. If anti-individualism about content is right, then in order to reap the metaphysical benefits of an account of mental causation cast in terms of supervenience and trope identities, then the anti-individualist must embrace a similar anti-individualism about causal powers. Without such an assumption, the causal powers of mental states cannot be identical to lower level causal powers; and without such a relation, the

causal powers of mental states would involve something of a mystery: How could it be that the intentional properties of mental states are causally relevant if they are completely distinct from the physical properties that we know are part of the causal order? Such a mystery is unacceptable in the cognitive sciences insofar as they intend to reject any sort of Cartesian substance dualism (something they should do since it is the causal mysteries of that doctrine that place it in opposition to a scientific account of mental causation).

It does not follow from the failure of psychological states to supervene on brain states (nor the failure to be trope identical with brain states) that causal relations involving mental states would be, from the point of view of rationality, undisciplined or chaotic. As Burge (1989a, p. 309) and Egan (1991, p. 181) point out, and as Fodor (1994) seems now to realize, so long as one's environmental relations remain stable, the fact that mental states do not supervene on neural states alone would not entail any sort of irrationality.[33] Thus, it is not impossible for the anti-individualist to frame regularities and generalizations for the purposes of explanation.

The problem with anti-individualism arises instead when we turn to the metaphysical interpretations of the taxonomies embedded in those regularities, generalizations, and explanations. If there are no causal powers for intentional properties to be identical with, then there is no account of mental causation. It could (in some sense) have turned out that mental states supervene partly on the environment; but if there is to be an account of mental causation that attends this anti-individualist possibility, then there must be causal powers that involve the environment in just the same way that the mental states do, causal powers on which the mental states supervene or with which they could be trope identical. The problem, we have seen, is that there is no reason to think that there are causal powers that are wide in the appropriate sense. In complaining, with some justification, about Fodor's H- and T-particle argument, Burge points out that "Fodor suggests that there is some mystery about failures of supervenience between psychological and physiological kinds, without explaining what it is". (Burge, 1989a, p. 309). We now have a way of articulating what that mystery might be. Anti-individualism asks us to countenance causal powers that are *unprecedented* in any scientific domain. This, I believe, is the sentiment Fodor was, or at least should have been, aiming to capture.[34]

As Burge, Wilson, Egan, McClamrock, and others have shown, however, some of Fodor's attempts to expose that difficulty are not successful. Individualism in psychology does not follow from superficial considerations of scientific taxonomy, but more plausibly from general considerations involving the relationship between scientific taxonomy and causal powers. CT and other tests for probabilistic relevance help us

to see the causal commitments underlying type-identification in the special sciences. Thus, even if we heed Burge's advice and put explanatory practice ahead of ontological considerations when thinking about causation and causal powers, we get no support for anti-individualism. It is the province of the lower level sciences to lay out the space of causal powers from which the causal powers of psychological states are derived. Burge's principle reasons for resisting CPN, and hence his principle reasons for resisting the repaired version of the Fodorian argument, are insufficient. If we grant (1)-(3) in this argument, as the anti-individualist must, then the individualistic conclusion is forthcoming. Once again, inherent difficulties in the anti-individualistic position motivate the search for a plausible individualism.

CHAPTER 6

COGNITIVE EXPLANATION

1. *Preliminary Remarks*

In the previous chapter I argued that there is no reason to think that causal powers are ever individuated relationally. Thus, to the extent that a psychology that embraces anti-individualism is committed to a relational individuation of causal powers in order to exploit the metaphysical virtues of supervenience, it faces difficulty. Looking to explanations in psychology to substantiate relational individuation of causal powers, however, is only one way in which the appeal to explanation might be taken to militate in favor of anti-individualism. In this chapter, I will address appeals to explanation that do not aim to show that causal powers are wide, but that nonetheless aim to establish that psychological states are to be individuated in terms of external states of affairs.

Under the banner of naturalism, anti-individualists appeal to explanations in the cognitive sciences in an effort to show that cognitive systems are full blooded members of the natural order. Seeing them this way, it is argued, forces upon us a reconception of cognitive systems. The traditional Cartesian separation between mind and world is seen as somehow antithetical to these naturalistic demands (McClammrock, 1995, is a clear example of this sort of thinking). I think this is a big mistake. In the first place, the metaphysical separation of mind and environment that is the heart of Cartesianism does not entail that cognitive phenomena fall outside the natural order. And secondly, the general line of argument that seeks to blur the distinction between mind and world, in my view, is either too strong or too weak. It is too strong if cognitive processes end up being *ubiquitous* in the environment; it is too weak if we are lead toward an *eliminativism* about the mental. In either case, there is nothing left that will distinguish we thinking organisms from the rest of the natural order? Cognitive processes can be seen to be world-involving in a way that is satisfying to anti-individualists, I believe, only when their distinctively cognitive features are ignored.

I will begin to argue for these claims by providing a preliminary characterization of what I take to be distinctive of the cognitive and psychological. I have already argued in Part I of this book that our

psychologies are characterized in part by certain epistemological facts; e.g., our knowledge of the contents of our thoughts is privileged while our knowledge of the environment is empirical in nature. Only internalism, I argued, can account for these facts. In this chapter, however, I will point to features that, in my own idiosyncratic way, I take to be characteristic of cognitive systems; they constitute some of the reasons why I think that cognitive states and processes supervene on an enclosed system such as is found in the heads of organisms. The main point that I will emphasizing throughout this chapter is that there are no reasons deriving from anti-individualism to doubt that cognitive states and processes supervene on internal states and processes alone. In section 4, my preliminary suggestions about what is cognitive should give at least some idea how an individualist would respond to various attempts by anti-individualists to show that the nature of explanation in the cognitive sciences forces upon us the view that cognitive states and processes not only should be individuated by appeal to the surrounding environment, but literally *include* that environment as well.

2. *What Are Cognitive Systems? Some Preliminary Thoughts*

We can begin to answer this question by recalling the familiar point that our only known examples of cognitive systems involve biological brains (which is not yet to presume that such systems are *limited* to biological brains, nor that non-biological systems might be discovered to think). In appealing to a notion of cognition in characterizing certain kinds of biological activity, we are (either implicitly or not) seeking to distinguish one way in which we and other thinking organisms are different from those organisms and non-organisms we consider to be non-thinking. It might be, for this reason, sound strategy to begin to think about cognitive systems by attending to the distinctive features of biological brains; in particular, we should attend to those features of brains in virtue of which they might be cognitive. This would not preclude there being cognitive systems that are very much unlike brains. But the burden is on those who would make such a claim to set forth identity conditions of their own in virtue of which such a system would be cognitive. I will identify four features of brains that I take to be crucial to their status as cognitive systems (this list does not pretend to be exhaustive). I will then argue that these features militate in favor of individualism.

(1) *The Character of Information Exchange.* A distinctive feature of brains that stands out immediately is the nature of information exchange; within a biological brain, information flows in a manner quite different from that between brain and the environment. This information flow, this cognition, is insulated from the environment by perception and action. In perception, information from the

environment is filtered through transducers and attentional mechanisms that regulate the amount and variety of information to which the brain, or a specific part of the brain, is exposed. On the output side, motor control structures restrict the ways in which the brain influences the environment. While there might be some modularity of cognitive processes (see, e.g., Fodor, 1983), the set of such processes is insulated from processes external to the nervous system in ways that they are not insulated from each other; at a minimum, the flow of information between modules does not require perception and action.

The skull and body act as boundries that prevent wholesale interruption of the internal information processing system. Because of this, the information processing system is allowed to pass into and out of myriad different environments without being altered in ways that break down the internal coherence. Organisms interact with their environments, to be sure, but there can be no denying the vastly different character of information flow within the brain from that between brain and environment. Of course, environments remain relatively stable as different organisms pass into and out of them; and developing infants change drastically as their environments remain relatively stable. But this only reinforces the fact that there is an important distinction to be drawn between the environments and the systems that develop, inhabit, and travel among those environments.

The significance of this insulation lies in the fact that it preserves a certain stability of the information and organization within the brain. Environments, for the organism, come and go, but the internal space retains a coherence that in the normal case is the deep basis for continuity of thought, memory, self-knowledge, and I would suppose also personal identity. The stability is more than just contingent or accidental; without it, it is hard to imagine anything like coherent cognition taking place. And one should not confuse this stability of information processing structures with the frenzied and often chaotic activity that goes on as these relatively stable structures interact with each other. The point is that the way these structures interact with each other remains relatively stable and constant, and isolated from the environment in ways that allow the *same* cognitive system to pass from environment to environment. There is, moreover, a certain inter-dependence of component processes. Our ability to reason, for example, requires memory. Some cognitive capacities can be lost as a result of lesions or trauma; but when a capacity is lost in this way, we recognize it to be the loss of a *cognitive* capacity. Too much such damage and the system as a whole ceases to function.

It is interesting to note, also, the unencumbered interaction between cognitive and conative processes in brains. Though we are here moving closer to questions about agency and personal identity, as

opposed to mere cognition, the fact that an individual's beliefs and desires conspire seamlessly to give rise to intentions reflects the deeply automatic nature of information flow within the nervous system. In this respect again, the brain is very much a unit isolated from its environment.

(2) *Intrinsic Intentionality.* More deeply still, perhaps, is the point that this insulated system is the only one we know of that is capable of *intrinsic* representation (or intrinsic intentionality -- see Searle, 1983; Maloney, 1990). It is a familiar point that the distinctive feature of cognition is representation.[1] This is not the claim that all cognitive processes are conscious. Rather, it is only the claim that cognitive processes must be representational, not as a result of being used by another system to represent something, but as a result of nothing more than the nature of the system.[2]

Cognitive processes will, of course, be sustained by mechanisms that transform representations without themselves involving mediating representations. In computational theories of reasoning, for example, the mechanisms that take the system from a representation of premises to a representation of a conclusion do not themselves involve representational stages; they instead mediate between stages of representation. Such mechanisms, I maintain, are not themselves cognitive, and in no way explain why the system is a cognitive system. They are, paradigmatically, mere implementational mechanisms. Thus, the fact that cognitive processes involve stages that are not representational (or not intrinsically representational) should not motivate one to maintain that external, non-representational processes might nevertheless be cognitive in virtue of their participation in a larger cognitive process (if there should be any). For it would remain the case that these external, non-representational processes are non-cognitive.

(3) *Point of View.* The idea of intrinsic representation is most likely related to the notion of subjectivity (see, e.g., Nagel, 1978). Having a point of view is something that seems intimately bound up with our brains. It may not be that we have, necessarily, the point of view *of* our brains; recall Dennett's (1978) fantasy in which one is detached from, and able to look at, one's brain. Nevertheless, without it, it is doubtful that we would have a point of view at all. Whatever sentiment there is for the view that computers in their present form cannot think, or at least do not think, probably derives from the belief that the computers have no point of view.

(4) *Loci of Control.* Another distinctive feature of biological brains is that they are the loci of control for organisms that have them. Both thought and action are directed and guided, as opposed to chaotic and aimless. Though the metaphysical nature of control is rather

murky, there is no denying that brains control the thought (or cognition) and behavior of organisms in ways that the environment simply does not. This is not to deny the causal influence of the environment on the cognition and behavior of organisms; I mean only to assert that the nature of internal control is quite different from the nature of external influence.

I do not for a moment take the aforementioned features to constitute a list of necessary and sufficient conditions on cognitive systems. I take them to be fairly pedestrian observations about the nature of biological cognition. That is, in at least the respects mentioned above, our token cognitive states are parts of systems that are, in a very clear sense, walled off from the environments in which they reside; they represent those environments from a particular and distinctive point of view, and they control navigation through those environments. That all this is true, I take it, is beyond dispute. And it is true in spite of the myriad ways in which we interact with our environments.

None of this is inconsistent with anti-individualism of the Burgean variety. Burge is concerned only with the individuation of mental states (and in particular, the individuation of mental states by appeal to their contents). Nevertheless, the idea that individual cognitive systems are separated and distinct from their environments in the ways suggested above has two important implications. First, if cognitive systems are isolated in this way, then it is not unreasonable to ask for arguments from Burgean anti-individualists to justify the claim that, in spite of this insulation, the token states inside the system should be individuated by appeal to factors outside the system. Thus, I take it that this isolationist picture of the brain, together with the presumption that the brain is crucially involved in cognition (and is the only thing we know of for sure that is engaged in cognition), foists upon the anti-individualist a burden, the burden to show that individuation must look outside the brain. Second, the isolationist picture places an even heavier burden on those who would like to claim not just that cognitive states and processes should be *individuated* by appeal to the environment, but that they literally *extend* into the environment. I will call this new variety of anti-individualism *systemic anti-individualism*, and I will examine how well its defenders discharge their burden in section 4 below. I hope to show that naturalistically motivated attempts by anti-individualists to appeal to scientific explanations to show that cognitive systems extend into the environment invariable offer no reason at all to abandon the individualistic picture I have been painting. Despite suggestions by McClammrock and others, I do not take appeals to, e.g., intrinsic intentionality, subjectivity, or the like to entail a repudiation of naturalism.

I will begin my case, however, by addressing the more conventional anti-individualism that we have been discussing throughout this book. I will examine certain explanatory virtues that, according to Wilson (1994a), attach only to an anti-individualistic psychology. This claim, I will argue, rests on a misconception of the nature of behavioral explanations, one that wrongly inflates the explanatory role of content. I will then consider various attempts (by, e.g., Tuomela, 1989; Haugeland, 1993; Wilson, 1994b; and Chalmers and Clark, in preparation, McClammrock, 1995) to show that an individualistic psychology arbitrarily draws the boundary of the mind at the cranium or skin. The over-arching theme of this chapter is that appeals to explanations in the cognitive sciences are often misused in the service of metaphysical (i.e., individuative) claims. One can reap all the explanatory gains purportedly accruing to anti-individualism while retaining individualism. Indeed, I shall argue, there are good reasons to do just that.

3. *Explanatory Virtues*

It is dangerous to draw metaphysical conclusions directly from scientific explanations. Explanations involve background assumptions that often go unnoticed, and, as we have already seen, these tend to obscure the metaphysical lessons to be learned from the appeal to explanation. I will argue in this section that the failure to recognize various kinds of background assumptions has led some to think, quite mistakenly, that there are explanatory virtues that only an anti-individualistic psychology can enjoy.

3.1 *Theoretical Appropriateness*

Wilson (1994a) describes an explanatory virtue, which he calls, somewhat cumbersomely, *theoretical appropriateness*. Good explanations are cast in terms that are, in a certain way, appropriate to their explananda. Explanations of the movements of planets that appeal to the behavior of their quantum-level constituents and the forces acting on them would be too unwieldy, and much less theoretically appropriate than a Newtonian explanation in terms of gravitational forces. Consider also Putnam's ([1973] 1981) example of a square peg being able to go through a slightly larger square hole, while it cannot go through a round hole with a diameter equal to the square hole's width. Though this can be explained either in terms of rigidity, size, and geometry, or microphysical properties of the objects involved, it is the former that is more easily understood, and hence more theoretically appropriate.

The virtue of theoretical appropriateness has not been lost on individualists. Many have recognized that if individualism is to be a

constraint on psychological taxonomy, then, if that taxonomy is to be theoretically appropriate to the explananda of psychology, there must be individualistic, or autonomous, descriptions of behavior (see, e.g., Stich, 1983). Autonomous behavioral descriptions abstract away from historical and relational properties of behavior, and focus only on intrinsic properties (e.g., movements), those that supervene on internal, physical states and movements of the behaver.

According to anti-individualists, this concession to the value of theoretically appropriate descriptions of behavior is a problem for an individualistic psychology. Wilson, for example, claims that autonomous behavioral descriptions do not fit "our conception of how behavior ought to be described for the purpose of psychological explanation." Part of the reason for this is that the individualist is proposing a revision of the explananda of psychology. This is evident in part from the fact that there are, Wilson maintains, difficulties in "fixing on clear cases of autonomous behavior." This, Wilson thinks, should be expected since an individualistic psychology "must differ in important ways from our commonsense psychology."

Others share these opinions. Tuomela (1989, p. 31), for example, thinks that an individualistic psychology would be reduced to explaining "...muscle contractions or efferent nerve impulses or something like that." Thus, because autonomous behavioral descriptions are not promising explananda for psychology, we must recognize that ordinary, non-autonomous, behavioral descriptions properly characterize the explananda for psychology. And because the individualist has conceded that individualistically individuated psychological states are not theoretically appropriate to ordinary behavioral descriptions, the individualist must concede that individualistically individuated states are not the proper explanantia in psychology. In this respect at least, individualism does not constrain taxonomy in psychology. And, to the extent that the taxonomy embedded in psychological explanations reveals the proper metaphysical individuation of psychological states, it follows that individualism is not true, at least for a large chunk of psychology.

3.2 *Contextual Background Assumptions in Behavioral Explanations*

In my view, the claim that behavioral explanations mandate non-autonomous descriptions of psychological explananda suggests no superior theoretical appropriateness for an anti-individualistic psychology. In the first place, it appears to be assumed that an individualistic psychology is obligated to embrace some sort of narrow content that is not about the external environment. But, as I have argued and suggested repeatedly, individualism is *not* obligated to accept narrow content. It may well be that mental states do represent the

external world, but that they have those intentional properties in virtue of some internal properties or tropes. Thus, there is nothing about individualism *per se*, as opposed to the more obscure notion of narrow content, that requires behavioral descriptions that are any different from the variety of description to which the anti-individualist avails herself. Any descriptions appropriate to an anti-individualistic psychology will be appropriate to an individualistic psychology that embraces broad instead of narrow content.

But even given this, the anti-individualist may argue that explanations of intentional behavior require appeal to more than what resides in the heads of individuals; thus, one is not only obligated to embrace broad content, one must admit that this broad content does not supervene on just what is in the head. Thus, the appeal to the nature of psychological explanation mandates not just broad (world-referring) content, but non-supervenient content as well.

I will argue, however, that this inference is fallacious. The fallacy consists in mistaken claims about what mental content is supposed to explain. It is true that ordinary descriptions of behavior are normally not autonomous descriptions. But explanatory appeals to content are obligated to explain behavior non-autonomously described only with the help of copious background assumptions about the *context* of contentful mental states and the behavior they help explain. On its own, I will argue, content explains only the intentional properties of behavior.

Many philosophers (e.g., Kim, 1982) have noticed that the reason that ordinary behavioral descriptions are not autonomous is that such descriptions normally also describe facts about the context of the behavior. But these contextual factors, *qua* contextual, are not what the appeal to content is supposed to explain. Thus, in providing explanations of behavior under such non-autonomous descriptions, the explanations must also appeal to certain non-contentful contextual facts *in addition to* the contents of psychological states. This point, or at least the significance of it, goes unnoticed, I believe, because it is often not noticed just what sorts of contextual assumptions are made in such explanations. Once these tacit contextual assumptions are made explicit, it will not be difficult to see that theoretical appropriateness does not militate in favor of anti-individualism.

There are at least three sorts of context, facts about which are normally assumed in explanations of intentional behavior: One sort of context, which we may call the *internal context*, is that provided by the body of the behaver. A belief-desire explanation of Biff's behavior assumes that once the belief and desire have conspired to give rise to an intention, say the intention to wave to Chip, the processes that are to effect this intention all run smoothly to the eventual waving behavior.

We can explain Biff's waving behavior by appealing to the content of Biff's beliefs, desires, and intentions only on the assumption that the internal processes that effect the action Biff intends are operating normally. If Biff's motor system were to encounter some sort of difficulty (e.g., muscle fatigue, motor neural inhibition, etc.), we would not blame Biff's *mental states* for the failure of his behavior. That is, if something were to go wrong with Biff's behavior "on the way to the world", so to speak, it would not be a problem that we would expect the appeal to Biff's mental states and their content to be able to handle.

A second sort of context, which we may call the *external context*, is that provided by the world in which Biff is behaving. The ordinary belief-desire explanation of Biff's behavior assumes that there is nothing abnormal inciting Biff's hand to move in a waving motion. We can explain Biff's waving behavior by appeal to Biff's mental states and their content only on the assumption that the circumstances external to Biff fall within a certain normal range. As with problems involving internal context, if Biff's environment were to fail to cooperate (e.g., something traps his hand behind his back, etc.), we would again not blame Biff's mental states for the failure of his behavior. That is, if something goes wrong "in the world", so to speak, it would not be a problem we would expect the appeal to Biff's mental states and their content to be able to handle. This point is even more dramatically evident in cases of so-called non-basic actions, those that have a large external contextual contribution. Hitting a home run in baseball, for example, requires more than a smoothly operating psychology; one's motor capabilities, the speed and placement of the pitch, the wind, the size of the park, and many other factors, non-contentful and therefore *non-psychological* factors, must conspire if the behavior is to come-off as he intends.

Another variety of external context involves *historical factors* in one's relation to the environment. As many philosophers have pointed out, one cannot sell a car that one does not own; if one is to sell a car, one must previously have stood in the ownership relation to that car. Were we to explain why Biff sold his car to Chip, we would have to assume such an historical context. If Biff *tries* to sell to Chip a car that he believes is his own, only to discover (what he previously did not know) that the car is in fact not his, the selling behavior would not be successful. While we might want an *additional* explanation to help us understand why Biff wrongly believed the car was his, the explanation for the failure of the selling behavior would not in the first instance cite any failing involving the beliefs, desires, and intentions that resulted in Biff's trying to sell the car. Given the belief that he stood in the ownership relation to the car, we may assume that there is nothing anomalous in Biff's decision procedure; the selling behavior was

unsuccessful for reasons quite independent of Biff's mental states and their contents.

Attention to these contextual factors is crucial if we are to see why the explanation of behaviors under non-autonomous description does not militate in favor of the theoretical appropriateness of anti-individualistic individuation in psychology. Take Wilson's (1994a, p. 64) example of latent learning among rats learning to navigate a maze; the specific way in which they do this is a matter of context, both internal and external, that is not relevant to what they learn. When the solid surface is replaced by water, rats are still able to negotiate the maze (provided they know already how to swim) because the rats learned something that was independent of running or swimming; one assumes that *already* knew how to run and swim. In appealing to the rats' learning, we explain their navigational understanding, not the particular way in which they implement that behavior, and not even whether that understanding results in successful completion of the behavior. (Each of these may offer *evidence* of the rats' navigational capacities; but the evidence, of course, does not *constitute* the capacity.) The change in external context (from solid surface to water) affects factors of the behavior that the learning was never supposed to explain; it was assumed from the start that the rats could run and swim. The change of context is of no relevance to what the rats have learned (the *content* of their learning, as it were), except insofar as it reveals to the experimenter that the rats have learned the layout of the maze rather than a particular behavioral implementation for getting through it. Had the maze been switched, after learning, to quick-sand, such that the rats could no longer navigate it at all, again nothing that is relevant to an explanation of the rats behavior by appeal to content would have changed.

The larger point here is that, for reasons not informed in any obvious way by metaphysical intuitions stemming from individualism, relatively uncontroversial analyses of behavioral explanations can tease out easily what is to be explained by appeal to mental states and their content, and what is to be explained in part by such appeals to contentful states, and in part by background contextual assumptions. To put it crudely, what the appeal to intentional states alone explains are further intentional facts; in this case, it is the *trying* (or the *intention*) that we expect the appeal to contentful states to explain, not whether the behavior is successful; for the success of behaviors is plainly dependent upon contextual factors that are quite independent of an individual's mental states and contents. Citing purely contentful factors in the explanation of a particular behavior under non-autonomous description is often taken to be adequate, but *only* if the (usually tacit)

contextual assumptions are made, and are in fact true (such that they go unnoticed).[3]

A great deal of sympathy with anti-individualism can be explained, and criticized, in light of our attention to the distinct roles of contentful factors and non-contentful background factors in explanations of behavior. Consider, for example, Tuomela's (1989, p. 34) example of well-trained and novice window openers. Tuomela claims that the autonomous description of window-opening behavior will include in its extension just the muscle movements involved in opening a window. But while the average person may have to, as we might say, think about his action of opening a window, a well-trained window opener may have window-opening as a basic skill, such that he need not pay any attention to it when he opens windows. In spite of their autonomously identical behavior, it is clear that while the layman's muscle movements are, in an obvious sense, intentional, the expert's are not; he does not even think about the movements he makes. Therefore, Tuomela says, there is no internal contribution to the production of full-fledged non-autonomous behavior that serves just as well as the psychological explanandum. Autonomous behavioral descriptions, even when connected to non-autonomous behavioral descriptions by contextual assumptions, do not constitute adequate (or theoretically appropriate) psychological explananda.

Tuomela's mistake, in my view, is that he does not recognize that appeals to intentional states alone explain, in this case, only the intention to open the window. The resulting behavior can be explained by the beliefs and desires that result in the intention to open the window, but only if, in addition to the usual external contextual assumptions, certain *internal* contextual assumptions are true as well. But, and this point is key, these assumptions may differ for different behavioral tokens. That is, the internal contextual assumptions needed to account for the layman's behavior may involve sub-routines that are *also* performed intentionally, while the same is presumably not true for the expert. But this does not affect the adequacy of the explanation of the original window-opening behavior in terms of the beliefs and desires that result in the intention to open the window. In both cases, let us imagine, the desire to ventilate the room, together with appropriate beliefs, explains the intention to open the window, and, provided the appropriate contextual assumptions are true, they will explain the window-opening behavior as well. We overlook this point when we ignore the fact that internal contextual assumptions are just as important to the explanation of intentional behavior as external contextual assumptions. Tuomela explicitly recognizes the external contextual assumptions, but never acknowledges the internal contextual assumptions.

Tuomela also ignores the general consequences of the fact that behavioral explanations typically involve contextual assumptions. To assume that there are no ropes tying Biff's arm to his side, or no force-field affecting Biff's movements, is *not* to make a *psychological* assumption, an assumption that involves Biff's intentional states; neither is it a psychologically relevant assumption that Biff's body is not damaged or otherwise out of order. These assumptions are not about Biff's mental states, yet they are the assumptions that account for aspects of the behavior that intentional states alone cannot account for. But, of course, these are precisely the sorts of assumptions made in the context of behavioral explanations. Tuomela is therefore quite right to claim that

> "...in the case of varying environments surely the environment has explanatory power, and arguably distal explanations will turn out to be indispensable...." (Tuomela, 1989, p. 33.)

But the conclusion to draw is not that these environmental factors have somehow (magically?) become psychological due to the fact that behavioral explanations make assumptions about them. Rather, the lesson to draw is that the psychological portion of behavioral explanations, the portion that appeals to the agent's intentional states, is intended only to account for what is left after the contextual assumptions have been taken account of. My suggestion, a suggestion that is both justified by careful attention to the nature of behavioral explanation and congenial to the individualist, is that the intentional component of behavioral explanations alone can account only for the *intention*, or the *attempt*, to act in the way that, if the contextual assumptions are correct, the individual does act.

If we speak loosely, we may say that the intentional component in behavioral explanations can account for even non-autonomously described behavior, but it can do so *only if* the contextual assumptions are in fact true; that is, given that the contextual assumptions are made, if they are true, then appeal to the individual's intentional states will predict (and, loosely speaking, explain) even non-autonomously described behavior. For example, if Biff's body is working properly and his hand is not tied to his side, then the fact that be believes that Chip is in view and wants to signal him will explain Biff's waving behavior. This non-autonomous behavioral description is theoretically appropriate to individualistically individuated psychological states, provided only that the internal and external contextual assumptions are true, and we recognize that behavioral explanations include these non-intentional assumptions.

Thus, there are two ways in which we might express the points about explanation that favor individualism: First, the intentional portion of behavioral explanations are merely aimed at explaining intentional phenomena (e.g., further beliefs or, in many cases, intentions or attempts to behave in ways that, if certain background conditions are met, one does in fact behave). Second, we may say that *so-called* intentional explanations are aimed at explaining behavior, where behaviors consist in movements under ordinary non-autonomous descriptions and perhaps other non-basic actions, but *only* with the help of non-intentional contextual assumptions of the sorts itemized above. It *is* true that ordinary behavioral descriptions are not autonomous; but it is *also* true that we do not expect the appeal to intentional content in ordinary explanations of behavior to account for those respects in which the behavior is not autonomous. In explaining behaviors under non-autonomous descriptions, we simply *assume,* as a part (a non-intentional part) of the explanation, certain facts about the various sorts of context in which behavior occurs.

When behaviors do not reach fruition, we may discover that one or more of these assumptions is mistaken. But, importantly, we do *not* take such discoveries to invalidate the intentional portion of the explanation; in making certain *contextual* assumptions, we do *not* take ourselves to be making *psychological* assumptions. Though I think this point is intuitive, it can be supported by the sorts of considerations raised in the previous section. All external and historical contexts, and some internal contexts, would by judged by those criteria to be non-cognitive. What we must notice, then, is that what are normally passed off as *psychological* explanations of behavior, are actually explanations of behavior that have *psychological and non-psychological* components; that is, explanations of non-autonomously described intentional behavior contain appeals to intentional states *and* appeals to non-intentional contexts. I will henceforth refer to these explanations as *behavioral explanations* so as not to suggest illicitly that everything appealed to in the explanation of intentional behavior is, of necessity, relevant to the individuation of intentional states. When we do take account of this, we can see that even if anti-individualists were right that Stich's autonomous behavioral descriptions do not fit our antecedent conception of how behavior is to be described for the purposes of psychological explanation, this is only because Stich, too, has failed to see what it is that we expect appeals to intentional states to explain. We do explain non-autonomous, intentional behaviors, it is just that we do not explain them by appeal to intentional states alone.[4]

3.3 Causal Depth

Wilson (1994a) advances another argument under the auspices of an aspect of explanatory power that he calls *causal depth.* Good explanations, he claims, have a certain generality. They appeal to states and entities that would still exist were things slightly different. And those entities would still cause in these nearby possible worlds what they have caused in the actual world. When explanations appeal to states or entities that would survive, and retain their causal efficacy, in nearby possible worlds, they are causally deep. Such explanations do not aim at generality merely for the sake of generality; after all, it is trivially easy to state generalizations that are *so* general as to be vacuous. Causally deep explanations are as general as they need to be, general enough to avoid being too immersed in minutia, but not so general as to wash over more fine-grained regularities.

Consider a wide conception of folk psychological states, which individuates those states by appeal to the environment, and a narrow counterpart theory that preserves the internal functional divisions of wide folk psychology, but instead individuates particular states in terms of their narrow content (see Fodor, 1987, ch. 2, on narrow content); call this 'narrow folk psychology'. Now suppose that we discover sub-personal differences in narrow folk psychology between creatures such as ourselves and other creatures whose behavior is as sophisticated and systematic as our own. Wilson maintains that this would not threaten the attribution of *ordinary* folk psychological (personal-level) states to them. There is, he claims, "a huge epistemological gap between investigations in sub-personal, cognitive psychology and the commitments of folk psychology." (Wilson, 1994a, p. 67.) There are no differences in the internal functionality captured by narrow folk psychology that could make a difference to the attribution of beliefs, desires, and the like.

Wilson's conception of folk psychology is, in my view, much too behavioristic. Imagine a creature who lacks some of the sophistication and systematicity of our behavior, say someone who has had a frontal lobotomy and is no longer capable of aggressive attitudes (aggression, I take it, is a perfectly ordinary category of folk psychological state). Why does this individual lack the capability of aggression? It is surely not *because* the individual lacks the behavior or behavioral dispositions normally associated with aggression. That is our *evidence* that the individual is no longer capable of aggression, but not the explanation of the deficit. Presumably it is because of some change in the individual's internal functionality, specifically, that functionality provided by the lobe that has been excised. The relevant behavior is not constitutive of the folk psychological state, but a symptom of it.

With the constitutive link between behavior and folk psychological state broken, we can imagine cases where the two come apart. First, it is possible in principle to have a folk psychological state *without* behaving in any way at all. It is therefore possible in principle to discover some internal functional description of an individual in virtue of which she exhibits the behavior that is the normal evidence of the presence of folk psychological states. Second, it is also possible in principle for an individual to display sophisticated and systematic behavior without having this sort of internal functionality. The 'replicants' in the film *Bladerunner*, or the robots in the film *Westworld* might be taken as (at least conceivable) examples of the relevant sort of creature. It is, at least, unclear whether replicants and robots have the beliefs, desires, and emotions posited by folk psychology (perhaps they are behaving only *as if* they had *real* beliefs and desires). Thus, it may be that folk psychological attributions *do* presuppose a normal internal functionality such that, where that functionality is lacking, so are the ordinary folk psychological states. At any rate, we have been given no compelling reason to think otherwise.

But Wilson's point is not that there fails to be some narrow functional description that underwrites the wide functional description, but that a heterogeneous mixture of internal functional descriptions can (and does?) underwrite the same wide functional description. Thus, increased generality, and therefore causal depth, can be bought by moving away from the narrow account and to the wide one.

In light of the possibility that narrow functionality may be functionality enough, we may question whether this quest for increased generality is well-advised. Wilson everywhere assumes that generality corresponds to causal depth, but it is obvious that some generalizations are *too* general. Consider the generalization that everyone who marries intentionally does so for the same reason, namely, because it is deemed better than the alternative. Obviously, this generalization, though surely plausible, washes over relevant differences in the causal/intentional explanation of the various instances of marrying behavior. There is *no* sense in which such a generalization picks out the deep causes of marrying behavior, despite the fact that it is more general than the specific intentional accounts that can be brought to bear in specific cases. This illustrates the point that some, even many, of the differences in internal functionality between two creatures will call for differences in the explanation of their behavior. It may be that explanation in terms of belief and desire are appropriate for creatures who differ arbitrarily in internal functionality, but *which* beliefs and desires are cited in the explanations of their behaviors are bound to differ as well. Thus, the

better explanation will be the one that takes account of these differences. Wide folk psychology may be inherently more general than narrow folk psychology, but this is sure to be a liability in many cases.[5]

Wilson's notion of causal depth looks to be rather unprincipled; just how much generality indicates the deep causes of behavior, and how much simply misses those causes?[6] Whatever the proper answer to this question, I think Wilson ultimately misses a very basic point. Internalism about folk psychology is the view that folk psychological states supervene on the internal states of an individual; the signature feature of mind-brain supervenience is that there is no difference in the mental without a neural (or neuro-functional) difference. That is, sameness of neuro-functional kind entails sameness of mental kind. But the possibilities that Wilson describes all concern the sameness of mental kind across *differences* in neuro-functional kind, and this is perfectly *consistent* with supervenience; the possibility of multiple realizability taught philosophers long ago that difference in neuro-functionality is not sufficient for differences in mental kind. Supervenience is theoretically attractive for precisely this reason. To be relevant, Wilson's claims must bear on whether neuro-functionally identical creatures could nonetheless differ mentally; i.e., whether neuro-functionally identical creatures could nonetheless merit different behavioral explanations. But nothing he says even bears on that claim; it is simply irrelevant to claims about supervenience that some differences in internal structure and function will not make a difference to the proper explanation of behavior. The salient fact, sufficient to establish some sort of local supervenience (or trope identity), and hence individualism, is that there is no difference in the mental causes of behavior without some difference in internal functionality. That Wilson fails even to address this fact renders his discussion of causal depth in folk psychology ineffective. There is every reason to believe that internalistic folk psychological explanations are as causally deep as they need to be. Therefore, causal depth is not a reason to favor an anti-individualistic construal of folk psychology.

Perhaps there is something to Wilson's point in application to explanations in the cognitive sciences. On first look, one might be skeptical that there is such application for the following reason. The dominant strategy of explanation in cognitive psychology appears naturally to support individualism. The strategy, what Haugeland (1981) calls 'systematic explanation' (Wilson calls it 'humuncularism'), first involves identifying some function that a system performs. We then postulate a series of functional mechanisms (what Haugeland calls 'intentional black boxes') to explain how the original function is carried out. We then perform the same move on the postulated mechanisms, until we eventually bottom out in hardware. How could this strategy

fail to be fully individualistic? All we ever do in providing such explanations of the capacities of organisms is look *inside* them.

Wilson resists this argument by pointing out that the description of a function may well be world-involving. Take the example of a face recognizer. There is no way to capture the generality of the function of a face recognizer without referring to faces. We can describe particular instantiations of face recognizers individualistically, of course, but no such narrow functional description will capture the fact that many internally different mechanisms might all share the function of face recognition. In "nearby possible worlds", we will have face recognizers that differ in internal functionality, and are only united by the "wide function" of recognizing faces. The reason for this is that "...face recognition is a modular capacity selected for the advantages that it confers [which]...would prevail even were the way it evolved to differ" (Wilson, 1994a, p. 70). Failing to make reference to the wide function of such devices therefore deprives the accompanying explanations of causal depth. Face recognition is an *essentially* wide function.

Wilson is here offering us a conception of psychology as a biological science, one that trades in adaptive functions, which, he claims, are wide. Some explanations in psychology are evolutionary explanations; and evolutionary explanations are patently wide. If individualism were to constrain psychology, it would have to constrain evolutionary biology as well. Since it does not constrain the latter, it cannot constrain the former, either.

I think Wilson's attempt to turn psychology into a branch of evolutionary biology is ill-considered, and there are other problems with his argument as well. I do not think that psychological explanation and taxonomy engenders any evolutionary commitments of the sort Wilson supposes, and that even if it did, the resulting evolutionary psychology could easily be shown to have no implications for cognitive psychology that conforms to the systematic explanatory paradigm outlined above.

First, it seems to me that, metaphysically speaking, the device is no more *essentially* a face recognizer than the world essentially contains faces. The fact of the matter is that the device does recognize faces in the actual world. But suppose it does this by responding to *anything* that has certain characteristics, formally distinctive of faces. It is *this* fact about it, together with the non-psychological fact that the actual world contains actual faces, that explains its success in recognizing faces in the actual world.[7] To borrow a distinction from our discussion in chapter 4, it may be that the function of the device in questions is only *identified* in terms of faces in the distal environment. When it comes to *individuating* the function of the device, it may be that the facts that make it true of the device that it plays the

psychological role that it plays are facts about the internal composition of individuals' cognitive systems; in this case, the device would play the same psychological role even if there were no faces in the environment, or no environment at all. Wilson must give us some reason for thinking the device is individuated in terms of external factors, rather than merely identified in that way.[8]

Moreover, one way to understand what it is for the device to be essentially a face recognizer is that it takes unvarnished representations of faces as inputs, and yields classificatory representations of faces as outputs. On this way of looking at the function of the device, its function is parasitic on the nature of the states that constitute its inputs; it would not be a face recognizer unless it trafficked in representations of faces. But then the question of whether the device is essentially a face recognizer or not depends on whether its inputs and outputs are really representations of faces. And if they are, then the question of relevance to the individualism debate, our old familiar individuative question, is this: What are the facts in virtue of which these states are representations of faces? If the facts that determine (metaphysically) the intentional properties of those representations are internal facts, then individualism remains true. If, however, it turned out that the set facts that determine the intentional properties include facts about the environment, then anti-individualism would turn out to be true. The important point to see, however, is that whether individualism or anti-individualism is true is *independent* of the devices function as a face recognizer. What matters is not *that* the device is a face recognizer, but *why* the device is a face recognizer. It is a face recognizer in virtue of the fact that deals with representations of faces, but it is an *open question* whether these inputs and outputs have their representational properties in virtue of internal or external facts. So, since the question that would settle the individualism dispute remains an open question no matter what the device's function is, questions about what function to *attribute* to the device are irrelevant to the metaphysical question of how the device is to be individuated.

Aside from this point, it is perfectly possible that the selective advantage face recognition confers on us (e.g., evading predators) would have been conferred on us by something else (who knows what -- heightened auditory sensitivity, perhaps), had the genetic basis of something else occurred in the gene pool. Wilson has given us no reason for thinking that this could not happen. One can imagine deploying the restriction to nearby possible worlds to argue that face recognizers are more widely applicable and hence causally deeper than their narrowly individuated counterparts. I worry, however, that these would be nearby simply in virtue of the fact that a face recognizing device is selected for in these worlds. Obviously, this would be a question-begging

strategy. So, there is no reason to think that the wide individuation of the device is more widely applicable and hence more causally deep.

Wilson's reasoning embraces an optimization strategy in interpreting evolutionary solutions to survival problems. The presumption seems to be that face recognizers are so valuable that, come what may genetically, we had to find ways to recognize faces. But recognizing faces is not the ultimate goal in survival, and face recognizers may simply be crude but suitable means for achieving that goal. Had things been even slightly different, we might never have needed to recognize faces. As with all evolutionary solutions, there is no reason whatever to suppose that it is the ideal solution from an engineering point of view. And if it is not, there is no reason to suppose that natural selection would have discovered this solution, come what may. Thus, there is no evolutionary reason for thinking that face recognizers and other widely individuated (or, to avoid begging the question against individualism, widely *described*) devices are more generally applicable across counterfactual changes. There is no reason to think that they are more causally deep.

The suggestion I wish to make at this point is that, for purposes of cognitive psychology conforming to the systematic decompositional strategy, there is no reason to look to evolutionary history in *individuating* the function of the mechanisms of cognition. Their functions can be *identified* by appeal to evolutionary history, but this history should be counted merely as *evidence* of the device's function; and that evidence is not constitutive of the device's functionality. The function of the device should instead be cast in terms of the role it plays in the system, where this could be given in terms of the interaction of contentful states that are individuated individualistically (rather than by appeal to their normal distal causes or targets). If this role changes, such that the functionality of the system deviates from what is statistically normal, the device's functional contribution will change accordingly. We may talk *as if* the device had a purpose or teleology, but I take this to be at most an expository convenience, a metaphorical importation from our own "design stance" (to borrow Dennett's terminology) toward the system. And, as I have argued, for purposes of metaphysical individuation, we must consider this functionality across a range of possible environments. By doing so, we might come to see that factors external to the system are irrelevant to the functional role of the mechanism in the system; the environment might be wholly simulated and that would not change our assessment of the functional role of the device within the system. Contingent features of the environment would therefore be irrelevant to the individuation of such a device.[9]

Appeal to evolutionary considerations might be methodologically sound, since such considerations might help us to

discover the functions certain mechanisms are performing; in fact, we may not be able to discover those functions without attending to the evolutionary development of the mechanisms. But this is an *methodological* point, not a *metaphysical* one. There is no reason at all to expect that the functions discovered in this way are *essentially* historically individuated.[10]

Even if (apparently contrary to fact) evolutionary factors could be brought to bear on matters psychological in something like the way Wilson envisages, I do not think they would have any impact on explanation and taxonomy in either folk psychology or a cognitive psychology that conforms to the systematic decompositional strategy. Selectional explanations account for the presence of certain traits in a population. Though Wilson would have us believe that evolutionary biology and so-called evolutionary psychology are in the business of explaining why creatures behave or function the way they do, this is misleading. As he later recognizes (p. 73), these enterprises do not explain the behavior of any individual creature; rather, they explain why certain traits, perhaps behavioral tendencies, characterize certain populations or species. In this capacity, they are best seen as providing accounts of Dretske's (1988, pp. 42-45) structuring causes (i.e., how the system got set up to produce one sort of behavior in a situation rather than another). But our original concern had been with the explanation of particular behaviors in particular circumstances; these are the so-called triggering causes of behavior. Though psychology is surely committed to the existence of structuring causes of intentional behavior, there is no reason to suppose that psychology must explain how those structuring causes came to be present. Even if there are wide accounts of the presence of traits or behavioral tendencies of population members, that provides no reason whatever to expect that there will be wide accounts of particular behaviors in particular circumstances.

This line of argument is not lost on Wilson; he calls it "concessive individualism" since it concedes wide explanations in evolutionary contexts, while denying them in strictly psychological contexts. He finds this position to be untenable for two reasons. First, he claims, it is implausible to suppose that psychology does not traffic in accounts of structuring causes. One reason for this, he asserts, is that issues of theoretical appropriateness demand wide taxonomies in psychology; if wide taxonomies are concerned with structuring causes, then it is implausible to suppose that psychology is concerned only with triggering causes. Of course, this reason has been refuted earlier, so I will not consider it any further.

Wilson also cites developmental psychology as an area of psychology that is clearly concerned with accounts that explain the development and presence of structuring causes (since it is concerned with

"the acquisition and change in an individual's cognitive structure over time..." (Wilson, 1994a, p. 73)). Even if this point were granted, there are several reasons to think that the range of application of the point is sharply and narrowly circumscribed. First, much developmental psychology is actually individualistic. Patterson (1988), for example, has argued cogently that accounts of semantic development are individualistic. When a child's understanding of a word deviates from that of the linguistic community, anti-individualists cannot appeal to the child's membership in a linguistic community to argue for a literal interpretation of that word; the only alternative is to ascribe to the child an understanding that is idiosyncratic, not dependent on her social relations, and hence individualistic. Also, Nelkin (in press, esp. ch. 9 and 11) has argued quite forcefully that appeal to developmental accounts of concepts *must* be narrow; no social or external factors can account for the acquisition of very primitive concepts on which the formation of others must rest. A substantial portion of developmental psychology, then, is narrow.

The second objection to concessive individualism is that there may well be developmental concerns that require cognitive structures to be identified by appeal to learning history or ontogenetic maturation. But this will mean simply that the *explananda* of psychology are sometimes temporally extended, not current states that are historically individuated (nor, of course, are they current states individuated in terms of historical relations to the environment, as in standard attempts to individuate widely). This is significant because there is nothing wide about temporal extents *per se*; there is no reason why individualism must be concerned with only momentary states. And when the explanandum for psychology becomes an historical progression, as it does for some developmental psychology, the relevant respect in which the taxonomy would be wide is whether it appeals to factors in the individual's environment; and there is no reason to suppose that is the case. Moreover, when the concern is with nothing more than how things have developed, there will be no forward-looking impact on the causal powers of the structures. Nothing having to do with the historical individuation of the structures will be relevant to their role in the future development of the system; place a 'twin' structure within a developing system and it will not alter the development of the system at all. Thus, even in those limited areas of cognitive psychology that concern structuring causes, there is no support for an argument against individualism.[11]

Wilson's second reason for thinking that the appeal to the structuring/triggering distinction will not militate in favor of narrow psychological taxonomies is the claim that triggering causes can themselves be individuated widely. Though Wilson provides no

support for this claim, we may assume that he takes his general appeal to causal depth and theoretical appropriateness to be substantiation enough. But this appeal, we have seen, carries no weight; the previous chapter developed a comprehensive case against the view that triggering causes can be individuated relationally.

I note, however, that Dretske (1992; see also 1993) has offered an argument that is advertised as showing that even triggering causes are individuated widely. He describes a device, Gizmo, that, when turned on, hums and smokes in a characteristically unproductive way. Why, he asks, does Gizmo behave that way. One answer, which Dretske does not favor, is that Gizmo behaves that way because it has been turned on and the current running through Gizmo activates certain motors in the characteristically unproductive way. This answer describes the triggering cause of the behavior, and does so in a patently individualistic way. The answer Dretske prefers, however, is that Gizmo behaves that way because (due perhaps to carelessness during construction) the blue wire has been soldered to terminal 16 instead of terminal 18. This description of the cause of Gizmo's behavior is patently anti-individualistic; it appeals to events outside of Gizmo, and in the past.

In my view, however, the anti-individualistic description can be, in Dretske's words, "screened off" by the individualistic description. If the individualistic description were not true of Gizmo, then, even if the anti-individualistic description were (i.e., even if it were constructed carelessly, with wires soldered (in the past) to the terminals as described above), Gizmo would not behave in its characteristic way, humming and smoking. What matters is only the way the wires are connected *now*, and how the current flows *now*. The anti-individualistic explanation, therefore, is not the real explanation; the real explanation, the one such that were it not true the behavior would not have occurred, is the individualistic explanation. Dretske is unmoved:

> "This strikes me as a desperate move, a move inspired by a preconceived (and theoretically motivated) idea of what a causal explanation *should* be like in order to qualify as scientific. It ends up rejecting as redundant (or superfluous, uninteresting, irrelevant) precisely the explanation that we are all looking for in situations of this type." (Dretske, 1992, p. 9.)

Desperate? Maybe. Preconceived? Yes. Theoretically motivated? I admit it. But, within the context of triggering causes, it is also defensible. The defense is this: The explanation Dretske favors is not an explanation of Gizmo's behavior, construed not as a process (as in Dretske, 1988, chs. 1 and 2), but construed as a sequence of movements

appropriately interpreted; *a fortiori*, it cannot be a *wide* explanation of Gizmo's behavior. It is, of course, an explanation of how Gizmo came to be wired in this way. And as such, it does appeal to facts in Gizmo's past and outside his boundries. But it should not surprise us that in explaining how Gizmo came to be constructed as it is, we must appeal to such facts; individualism is not the view that Gizmo is *sui generis*! It is simply the view that the triggering causes of Gizmo's behavior are to be found inside Gizmo, not in Gizmo's history, or history of relations to his environment.

Dretske's point shows up the fact that sometimes when we appear to be asking for an explanation of a thing's behavior, we are actually asking for an account of how the thing came to be as it currently is; we already know what causes the thing to behave in this way, and now we want to know how those causes got there. There is nothing superfluous and redundant in this explanation, but only when we realize what it is an explanation *for*. Once we realize what it is an explanation for, however, we also see that it is not an interestingly anti-individualistic triggering explanation for, say, Gizmo's behavior; it is not an explanation of Gizmo's *behavior* at all. Therefore, nothing favoring the anti-individualism of triggering causes of behavior follows from Dretske's considerations.[12]

Thus, it seems to me, there is no good reason stemming from the analysis of behavioral explanations to think that states external to the individual are in any way themselves psychological states, or are in any way relevant to the individuation of psychological states. Since my posture in this chapter is essentially defensive (I aim to show only that there are no reasons for an individualist to feel threatened by appeals to explanation in a scientific psychology), this is enough for now.

4. *Systemic Anti-individualism*

In the interest of explanatory power, a nest of recent arguments have sought to show that there are paradigmatically cognitive processes that are impervious to the physical boundries that separate brains from the bodies to which they are attached and the environments in which they are embedded.[13] According to such arguments, to restrict one's attention arbitrarily to internal affairs is to miss phenomena that stand in need of characterization and explanation. The aim of these arguments is to establish a sort of anti-individualism, what I have called *systemic anti-individualism*, on which there is, from a psychological point of view, no principled distinction between much of what goes on inside the brain of an individual and what goes on in her body and environment.

4.1 Feedback and Learning

I begin this section with a somewhat mild argument that conforms to this general strategy. Tuomela (1989) makes the familiar claim that the fate of individualism is an empirical issue, to be settled by appeal to 'psychological' explanations of the highest quality:

> "My general claim is that it is reasonable to go by causal laws, so to speak; and these laws will probably dictate against [individualism]. ...[B]oth psychological explananda and their (best) explanantia would presumably typically be noninternal." (Tuomela, 1989, p. 37.)

Tuomela criticizes individualism on the grounds that it cannot account for behavior that involves adjustments based on feedback or learning. I believe that this criticism flows from misunderstandings exposed in section 3, as well as a deep misunderstanding about the nature of individualism itself.

Tuomela's argument is, in part, that

> "...in cases...where the organism gets feedback from its actions upon the environment and corrects its mistakes, and learns and directs its future actions on the basis of its feedback-based interaction with the environment, [individualism] is in trouble. For here the relevant proximal explanatory brain states are...*causally dependent* on the external world.... What is more, the proximal explanatory brain states corresponding to each action token in the series are probably going to be rather disparate...[reflecting the need for]...a kind of *distal common-cause* explanation of the disparate proximal explanatory states." (Tuomela, 1989, p. 39-40.)

Since the appeal to a distal common cause is required for a full explanation, and the distal common cause will, *qua* distal, be external to the system, there is no escaping the need for appeal to external factors in explaining feedback-based and learning-based behaviors.

One mistake in Tuomela's argument is the aforementioned failure to understand what are the proper explananda for explanations appealing to the contents of mental states. In discussing the example of hammering a nail, which requires adjustment based on feedback as the nail goes deeper into the wood, he supposes that the individualist could only offer explanations of "...the movements of the body involved in the various hammerings." But we have seen that this is simply false. The proper explanandum for an explanation in terms of representational content is, in the first instance, the intention or attempt to hammer in

the nail. To the extent that this is done by means of repeated hammerings guided by feedback, the various hammerings are mere details of the implementation of the intention to hammer in the nail, what I have been calling internal contextual factors. This does not prevent the individualist from explaining the hammering of the nail, of course, it is just that the explanation will not be an entirely content-based explanation; it will appeal to contextual factors as well.

A deeper mistake lies in what appears to be a familiar misunderstanding, on Tuomela's part, of individualistic psychology. Tuomela early on (p. 24) discusses Fodor's formality condition, and notes, correctly, that for representations in a computational system the formality condition "...requires syntactic mirroring of semantic content...." But by the time Tuomela gets around to criticizing individualism founded on something like the formality condition (p. 37), the mirroring in question, which is a constraint on the *encoding* of semantic contents in syntactic states, seems to be confused with the *representation* of the environment by the semantically endowed syntactic items (i.e., the internal representations). Tuomela speaks of

> "...the internal states that the [individualist] assumes to mirror or reflect external happenings..." (Tuomela, 1989, p. 37).

Given this, he goes on to say:

> "The mirroring in question may be distorted. It is supposed to reflect the organism's point of view, so to speak. When the organism makes perceptual mistakes, for instance, the mirroring is of course distorted. but -- and here we have a criticism against the Explanatory Thesis relying on mirroring -- the best explanation for the organism's behavior will have to account both for its successes and its failures, and this may require the explanation to employ noninternal descriptions going beyond what mirroring gives us." (Tuomela, 1989, p. 37, fn. 5.)

If this were the extent of the confusion, I would be inclined to say that Tuomela has simply chosen his words unwisely, using mirroring in two senses, one to explicate the encoding relation (i.e., how intentional properties are encoded by the formal properties of a system) and another to explicate the representation relation (i.e., how intentional properties reflect what they are about). But he attributes the familiar commitment to the individualist that they cannot "employ noninternal descriptions", on the basis of the fact that the individualist endorses some sort of "mirroring". I am not sure what he means by this, but he seems to be saying that individualists cannot appeal to representations

whose contents are given by descriptions of the external world. Indeed, he says later (p. 39) that individualists would have to explain feedback action sequences "...without full reference (that is, reference in noninternal terms concerning the agent-external world) to the causal contribution of the world." This sure seems to be saying that individualists are not entitled to talk of representations of the world.

But, I have argued, there is no such restriction on individualism. Only a narrow (non-referential, non-truth-conditional) account of content, or perhaps Stich's (1983) syntactic theory of mind or Churchland's (1981) eliminative materialism, could be correctly saddled with such a restriction. Individualism is identical to none of these views (even if these views are individualist in spirit). The more conventional form of individualism, the form of individualism that I am defending, is perfectly happy with representations whose contents are given in terms of descriptions of the world. The only constraint is that the representational (syntactic and semantic) properties must supervene on (or be trope identical to) internal states of the organism. The mirroring in Fodor's formality condition is *this* sort of mirroring, a mirroring in syntax of intentional content, not the extent to which representational contents (or syntax) mirror the outside world. To criticize individualism for a failing attaching only to narrow content views, or Stich and Churchland, would be to throw the baby out with the bath water.

Even this misinterpretation of individualism would not be worth mentioning if it did not have larger consequences. In the first instance, since individualism is able to traffic in representations of the external world, there is no reason to suppose that individualism cannot provide proper explanations of feedback action sequences such as hammering. The account broached above should suffice here. But even if we suppose that there are content-based explanations of the various token hammerings involved in the sequence, we do not get an argument against individualism.

On the usual sort of internal and external contextual assumptions, each instance of feedback information will be recorded by an *efference copy*, which is compared to a goal state. The efference copy is a low-level representation of the current external situation, in this case, the extent to which the nail is hammered-in. The goal state, in this case, is a fully hammered-in nail, which is presumably represented in the intention under whose governance the action is proceeding. As information is fed back to update the efference copy, the system will detect the amount of variance between the current external state and the goal state, and adjust any parameters necessary to make sure that subsequent activity will bring the efference copy into agreement with

the goal state. When the goal state is reached (that is, when the intention is satisfied), the action ceases.[14]

What is key in all of this is that both the efference copy and the representation of the goal state are *internal* states of the system. And there is nothing particularly disparate about these internal states; they are simply internal states that *represent* external (or distal) events. I maintain that, if the appropriate contextual assumptions are made (see above), the proper explanation for the hammering sequence can therefore be provided in terms of an individual's internal states. That is, it is in virtue of the individual's internal representations of external events that she behaves (i.e., hammers) the way she does. Indeed, we can even show that the external contextual assumptions are not necessary. Imagine an elaborate simulation of the external context that provides to the individual's various sensory apparatus (including proprioceptive apparatus) exactly the same stimulations that that apparatus would receive if the external contextual assumptions were true. The sensory apparatus, and the individual herself, would make no difference whatever between the two cases. From the individual's point of view, exactly nothing appears to differ across the actual and simulated environment. External events, then, should not necessarily be implicated in a content-based explanation of this behavior; the intentional machinations would be no different if the external events were entirely absent. Thus, when Tuomela claims that individualism is in trouble due to the fact that external states of affairs are necessarily implicated in the explanation of behavior, he is simply mistaken.

It is this mistake that underwrites Tuomela's (1989, p. 40) ill-conceived appeal to Reichenbach's principle of the common cause. Tuomela considers the example of a person walking toward a tree. His suggestion appears to be that the tree itself somehow obviates the need to refer to internal states in explaining the person's behavior. The tree is the common cause of the various perceptual representations of the tree tokened by the person as she walks toward the tree. As a preliminary point, note that Tuomela here appears to be concerned with something other than *why* the person walks toward the tree at all (obviously the tree does not explain why the person walks toward the tree), but rather *how* the person is able to walk toward the tree (or why the person walks in this direction rather than another). His claim seems to be that an explanation that cites the *tree*, rather than the various *perceptual representations* of the tree, is the best explanation of how the person is able to walk treeward; given the desire to walk treeward, the person is able to do so because the tree remains a common perceptual influence guiding her walking. The individualist cannot simply appeal to the need for internal representations of the tree, Tuomela says, because, barring reference to the tree as the common cause, it will only be an

accident if these various internal representations are nomologically unified. To achieve the causally deepest explanation, we must make reference to the tree as the common cause of the representations.

Tuomela's argument here appears to miss the point, noted above, that individualists can appeal to representations of distal objects without sacrificing the individualistic determination of content that defines their position. What we must make essential reference to in explaining the treewardness of the person's walking is the *representation(s)* of the tree, not the tree itself. Provided that the representation supervenes on (or is trope identical to) internal states of affairs, the individualist can exploit the fact that the various representations are all of a single tree in providing an individualistic explanation of the person's treeward walking. To see this, consider again a simulation thought experiment: Simulate all of the external stimulations that the person receives, and her behavior, as well as the mental states from which it results, will be indistinguishable from the actual case.[15] The individualist, armed with representations of distal events, can provide a perfectly coherent explanation of the person's behavior, one that is not parasitic on an anti-individualist account.

It may be objected, of course, that in helping myself to an individualistic view of broad content (i.e., content that is about the external world yet supervenient on, or trope identical with, internal properties of individuals), I am simply assuming what I should be proving.[16] That is, there is a common presumption that individualism is true only if there is some workable notion of narrow content. Burge and others have shown, have they not, that ordinary content-clauses describe contents that do not supervene on internal states, and that they fail to supervene because they are about contingent features of the external environment. The only content that could supervene locally, then, is content that is not about the external environment. But, of course, as Adams (1992), Lepore and Loewer (1986), and others have shown, even the purported descriptions offered as the narrow content of mental states are actually wide, in which case it would appear that individualism cannot have what it needs. And now, as Tuomela has shown, even if there were some notion of narrow content, it is explanatorily inadequate.

I will let my remarks in chapter 1 stand as an attempt to say how ordinary mental contents could be about the external world while remaining locally supervenient. The issue here, however, is about the contents appealed to in the service of scientific explanations of behavior. I have argued in chapter 4 that while there may be a common presumption that ordinary content is broad and not locally supervenient, there is no warrant for a similar presumption in the scientific context. This was especially clear in the case of visual content, which may

supervene on, or be trope identical to, physical properties of visual states. If, however, Tuomela is assuming that the common presumption from folk psychology carries over to the scientific explanation of behavior, then he is in danger of assuming what it is he should be proving.

Let me grant that narrow content, content that makes no reference to anything external to the subject, is explanatorily inadequate. If that is all that Tuomela is attempting to show, then we have no disagreement. But some of the quotations presented above suggest that Tuomela would also like to claim that content that is about the external world cannot be token identical to anything internal to the system in question. It is this claim that I maintain is not defended. If Tuomela is assuming that this claim is true without any support, then he has simply not made any case against individualism; for individualism may be true in spite of Tuomela's arguments if it should turn out that broad, world-referring content can be locally supervenient. If, however, he is assuming that this claim is true because he assumes that Burge's arguments in the context of folk psychology (that contents that are about the external world are not locally supervenient) also apply in the context of scientific psychology, then his argument does not represent a new argument against individualism; he has simply helped himself to the old argument. What he has shown, in other words, is that if anti-individualism is right about contents in scientific psychology (and not just folk psychology), then even scientific psychology is anti-individualistic. But who would question this? The real issue is whether anti-individualism *is* right about the contents appealed to in scientific psychology. For if anti-individualism is not right about contents in scientific psychology, then Tuomela cannot resist the defense of individualism presented above. If Tuomela requires Burge's arguments to carry over to a scientific psychology, then, in a sense, his arguments are not necessary. In either case, they pose no new threat to individualism.

4.2 Mind Embedded and Embodied

I would now like to move toward a more robust form of systemic anti-individualism. I will begin with the provocative work of Haugeland (1993), who attempts ingeniously to expose an "intimacy" in the relation between mind, body, and environment. By this he means "...a kind of *commingling* or *integralness* of mind, body and world...[that] undermine[s] their very distinctness" (Haugeland 1993, 2). The challenge, as Haugeland sees it, is to make this position intelligible; for once it is understood, its plausibility will be manifest. I will use Haugeland's case to introduce a general line of thinking, the banner of

which has been taken up by others (e.g., Wilson, Clark and Chalmers, McClammrock) whose work will be considered in the following section.

Haugeland's case is clever. He begins with a seductive example owing to (Simon 1969/82) involving an ant who travels across an irregular beach, the ever-changing surface of which is shaped by the random action of waves and wind. The ant's trajectory across the beach is a complex sequence of turns and weaves -- much too complex to be the product of simple ant-cognition. It is, in one clear sense, the beach's complexity that is responsible for the complexity in the ant's path. Perhaps, says Simon in a move echoed by Haugeland, the same can be said for people.

There is a lesson in this realization; or maybe two.

> "On the one hand, one might heave a sigh of scientific relief: understanding people as behaving systems is going to be easier than we thought, because so much of the apparent complexity in their behavior is due to factors external to them, and hence external to our problem. On the other hand, one might see the problem itself as transformed: since the relevant complexity in the observed behavior depends on so much more than the behaving system itself, the investigation cannot be restricted to that system alone, but must extend to some larger structure of which it is only a fraction" (Haugeland 1993, 4).

Haugeland, clearly, is pushing for the latter option; and maybe he has a point. He develops his case with the help of some insights from Simon's version of systems analysis. In particular, he wants to mark this distinction: interactions *between* components in a system vs. interactions *within* a component in a system. Interactions between systems are "narrow bandwidth" interactions; they are effected through a sort of bottleneck, such that what happens inside the component is well-insulated from what happens outside the component, except at the limited focus of the interface. Electrical components of a TV, for example, have narrow bandwidth interfaces. Interactions within a component, in contrast, are "high bandwidth" interactions; here the interface is not bottlenecked in any appreciable way. A random partitioning of the guts of a TV, one that doesn't respect the usual component divisions, is high bandwidth; and for that reason, it is not a scientifically interesting division. Indeed, the interesting divisions need not be corporeal or spatial. Social structures, particularly those that are organized around information flow, can be divided into components by appealing to intensity of interaction, regardless of spatial propinquity. Telemarketers working from the same room, for

example, may never communicate with each other; their more interesting interfaces involve their bosses and clients.

Armed with these analytical insights, Haugeland invites us to consider several examples of manifestly intelligent behavior that involve high bandwidth interfacing between the behaver and its environment. The first, of course, is the ant and its interface with the beach. Had the ant navigated by means of an internal map of the beach, then it would enjoy only narrow bandwidth interfacing with the beach (through the soles of its feet). But the fact of the matter is that the ant has no such internal map, and participates in a free exchange of information with the beach as its trek progresses; "...the ant and beach must be regarded more as an integrated unit than as a pair of distinct components" (Haugeland 1993, 12).

A related example comes out of the work of Brooks (1991), who has built an insect-like robot that navigates the MIT AI laboratory without the aid of a symbolic cognitive architecture. Instead, it has "layers" of largely independent behavioral units connecting sensation to action. Several layers are 'superimposed' and coordinated so that they don't interfere with each other in ways that damage performance. The most basic layers involve object avoidance, while others involve, e.g., picking up soda cans around the lab. According to the criteria for system decomposition, the most salient divisions are not between behaver and environment, but between the different layers. There is, in other words, high bandwidth interaction within layers, including the environment (which Brooks calls "its own best model"), and only narrow bandwidth interfacing between layers. As with the ant, it is not at all clear why the object of psychological inquiry is the behaver *per se* when the most interesting divisions do not include any such concept.

Haugeland also cites, with approval, Gibson's (e.g., 1979) ecological approach to psychology, and Dreyfus' (1972/79/92) persistent criticisms of classically oriented AI. Gibson's talk of "affordances" is well-suited to Haugeland's assertion that the behaver(perceiver)/environment interface is high bandwidth. Perceptual systems include not just the perceiver, but the perceived, and the relation between them is one of mere "pick-up". Affordances *unify* the perceiver and the perceived by essentially involving both. Dreyfus' well-known complaints about classical symbol systems do for mind/body interaction what Gibson has done for mind/environment interaction. The 'intensity of interaction' criterion calls into question the separateness of mind and body for cases Dreyfus presents. Performing any complex task even reasonably well requires a vast amount of information exchange between the brain and the body. There is very little sense, Haugeland thinks, to the notion of an instruction guiding

complex behavior. There is instead a high bandwidth interface between brain and body something like that of ant and beach.

Lest we think that all of the above is plausible only in the case of simple organisms, or at least simple tasks, Haugeland asks us to consider knowing the way to San Jose as an example that (at least for those who live there) hits quite close to home. A signature feature of intelligence, one almost entirely lacking in the cases presented above, is the ability to deal with what is absent. San Jose (for most of us) is absent, and knowing the way to San Jose involves the capacity to deal with what is absent. It is a central commitment of Cartesian cognitivism that such a capacity must involve internal representations of what is not before us; how else can we hope to direct ourselves coherently? Haugeland's suggestion, building on lessons drawn from Simon, Brooks, Gibson, and Dreyfus, is that the world takes care of much of the task. Knowing the way to San Jose is knowing what road to take, and, as it were, letting the road do the rest. The relation between a driver and the road taken is strikingly like that of the ant and the beach. High bandwidth interaction such as this is the mark of *no* (relevant) distinction; the road, no less than the driver, is an essential part of an intelligent system.

Haugeland has made it intelligible, even plausible, to suppose that the seat of intelligence, the mind, abides in the public world. If the universe of discourse for psychology is the mind, then the science of psychology must expand its domain beyond current boundries. Psychologists must reconceive their task fundamentally , and explore the myriad ways in which what is inside the head interfaces with what is outside the head. Or not.

4.3 Against Embeddedness

Haugeland's case depends (by Haugeland's own admission) on shifting the sort of problem it is the business of psychology to address. In a passage quoted above, Haugeland invites us to reappraise the domain of psychology on the grounds that there is an interesting problem to be addressed that is not currently being addressed by psychologists. The brain is still dishing-up output, but, sometimes at least, funny things happen on the way to the world; the complexity of behavior owes much to the complexity of the environment, and its relation to the organism. It is hard to see, however, why that would motivate a shift in the focus of psychology. Consider the ant and the beach. Haugeland asserts that "...what we want to understand in the first place is the ant's *path*" (Haugeland 1993, 11). But this is debatable. An ant psychologist may have a passing interest in the complexity of the beach, and even how that influences the ant's path, but what she would really want to understand, I wager, is the *ant's* contribution to that path. And what a

human psychologist is interested in is not the close coupling between the driver and the road, but what the driver brings to that coupling. It is quite uncertain, therefore, that Haugeland has given us compelling reasons to expand the universe of discourse for psychology. That there are other interesting questions to ask, even, perhaps, more interesting questions, does not eliminate questions of paradigmatically psychological interest.

But, Haugeland will be quick to point out, if we grant that the ant and the beach, or the driver and the road, enjoy high bandwidth interfacing, the onus is on us to articulate the criteria according to which we can make a (nonarbitrary) distinction between mind and world. There are at least two (related) avenues we might explore preliminarily, both of which suggest that the relevant bandwidth of an interface is, at best, insufficient to delineate component boundries in a system in general, and the scope of psychology in particular. The first can be introduced in the context of Brooks' layers of behavioral units. Even on the assumption that Brooks' subsumption architecture reveals a kernel of truth about the organization of behavior, it does not do the work Haugeland needs of it.[17] Any intelligent creature is involved in many behaviors, simple and complex, extending over many environments. But, to recall the first of distinctive features of cognitive systems mentioned at the outset of this chapter, while the environments may come and go, the organism (at least ideally) does not. One clear point of contrast between the inner and the outer, then, is that the inner remains (relatively) stable across significant changes in the outer. The inner/outer dichotomy is just not the illusory and arbitrary dichotomy that Haugeland makes it out to be.

This is all the more striking when we consider the case of the driver who does a good deal more than drive to San Jose. She may also hold down a job, go to dinner with friends, play tennis, study financial portfolios and any number of other things. Each of these may in some way involve high bandwidth interfaces with the relevant environments, but it is only *she* who enjoys high bandwidth interfaces with all the relevant environments. The point is that the object of psychological inquiry is determined by more than a single behavioral episode. A large part of AI, both classical and connectionist, concerns the exploration of models that generalize beyond a limited behavioral domain; we want to figure out how a single system could do so many different things. High bandwidth interfacing may be *necessary* for intracomponent interaction, but there is no reason to expect that it is *sufficient.*

The recognition of the intersection of different behaviors dovetails with a second point that tells against Haugeland's attempt to broaden the scope of psychology. Recall from the third distinctive

cognitive feature mentioned at the outset of the chapter that minds are typically viewed as loci of control. The insides of the ant, Brooks' robot, and the driver control their respective behaviors in ways that the relevant environments do not. Organisms do not determine the nature of their actions independently of the environment (a brick wall is, after all, a *brick wall*); but they do determine their actions in a way that the environment does not, namely, by deciding or choosing or at least being motivated to take a particular course of action. And while it may be difficult to articulate just what the relevant differences are between the inner and the outer (what makes a choice a choice?), there can be no doubt that they are real. Again, the presence of a high bandwidth interface may well be a necessary feature of intrasystemic interaction; but it seems manifestly insufficient.

4.4. Other Systemic Anti-Individualists

The points that I have so far utilized seem to me also to be successful in warding off another sort of argument for systemic anti-individualism. Chalmers and Clark (C&C, in preparation) provide several examples of cognitive processes that they believe are necessarily world-involving. One such example involves three cases. In the first case, a subject is seated before a computer screen and is asked to perform a task that requires her to look at a two-dimensional figure on the screen and rotate it mentally. In the second case, the subject has the choice of rotating the figure mentally, or physically by pressing a button on the computer; the physical method is much faster. In the third case, we imagine the subject with a neural implant that can rotate the figure physically on the screen, and the subject has to choose whether to do it the old-fashioned way, or the new hi-tech way.

C&C claim the mental rotation procedure of the first task is paradigmatically cognitive. They then claim that if each of the procedures in the third case involve cognitive processes, then there is no reason not to allow that each of the procedures in the second case are cognitive. But in that second case, the physical rotation of the object on the screen is accomplished by an external process of pressing a button. If they are right that that is part of a cognitive process, then they are right to claim that cognitive processes, and hence cognitive systems, extend into the physical environment, and are not necessarily contained within our heads.

C&C make similar remarks about another example involving an Alzheimer's disease patient they name Otto. Otto's memory has degraded to the point where he remembers virtually nothing, and instead enters important information (about, e.g., locations of things, names of people, relationships, etc.) into a notebook that he carries with him. Whenever he realizes that he is in need of information that

he cannot recall, he consults his notebook. If Otto wants to go to the museum, for example, he looks in his notebook, sees that it is on 53rd Street, and heads to 53rd Street.

C&C compare Otto's situation to that of a normal individual, Inga. When Inga decides to go to the museum, she recalls that it is on 53rd Street and goes there. It is natural to say, in this case, that Inga believes that the museum is on 53rd Street. Moreover, she believed this even before she consulted her memory; it is a standing belief of hers. C&C claim, however, that the same can be said for Otto. He, too, believes that the museum is on 53rd Street; given his desire to go to the museum, he goes to 53rd Street because that is where he believes the museum is located. But if Inga believed that the museum was on 53rd Street before she consulted her memory, why should we not say that Otto believes the museum is on 53rd Street before he consults his notebook. Otto's notebook, after all, plays in his cognitive economy the same role that Inga's memory plays in hers.

These arguments differ somewhat from Haugeland's insofar as they do not rely on bandwidth to establish the inclusion of the environmental in cognitive processes. Rather, they exploit the similarity between a paradigmatically cognitive task that happens to occur internally in one case, and another example of that task that in another case involves the environment essentially. It seems to me, however, that the same points that undercut Haugeland's argument can be pressed into service here as well.

In the first place, there is a relevant difference between the cases that C&C raise in that the world-involving cases involve perception and action. For this reason alone, it hardly seems obligatory for us to treat the explanations of the two kinds of cases as equivalent. In fact, it would seem that they are importantly different for precisely that reason. In the world-involving cases, the subjects have to *act* in a way that demands of them that they *perceive* their environment. There is a non-trivial story to tell in accounting for this aspect of their performances of the tasks. But there is not the slightest reason to think that, because the performance in the two kinds of case have the same result, they both result from *just* cognitive processes, let alone the *same* cognitive processes. The very fact that the results are achieved in such remarkably different ways suggests that the explanation for one should be quite different from the explanation for the other. In the Otto case, for example, there is the fact that Otto has to remember to look at his notebook while Inga does not have to remember to remember. Otto then has to look at his notebook, while Inga has to look at nothing. We are tempted to applaud Otto for his ability to cope with his illness; we have no such temptation with respect to Inga. There are, then, myriad

differences in the way we would explain the behavior of each. Similar remarks apply to the C&C's first example.

Second, whereas the processes in the head seem to traffic in representations that are intrinsically intentional, there is little reason to suppose that the same could be said about those representations external to the subject. This is not to presume, in a question-begging fashion, that the representational tokens in the head are individuated in terms of contents that supervene solely on internal factors. That dispute, the dispute over whether the content of representations supervenes locally, is a separate dispute involving separate arguments. If representational content is world-involving, then of course individualism is false. But the arguments presented throughout this book have been designed to undercut any support there might be for this claim. If those arguments are correct, then representational content is not world-involving. And if it is not, and the only intrinsic representation occurring is occurring inside the head, then there would seem to be little reason to regard the external portions of the processing to which C&C appeal as cognitive in any interesting way. Events external to the subject might facilitate the eventual success of the processes occurring internally, but that in no way suggests that these external events are therefore cognitive in nature.

Third, there can be no question that the locus of computational or cognitive control resides inside the head of the subject. In both sorts of example, the decision to engage in the task, the drive to complete it, and the way it is to be carried out all involve internal processes in a way quite distinct from the way external processes are involved. If this feature is indeed the mark of a truly cognitive system, then it is a mark by means of which the external processes C&C point to can be excluded.

Wilson (1994b) follows a similar line of thinking in arguing that computational processes extend out into the world. To take a simple example, we use pencils, papers, and the like to store information at intermediate stages in the process of, say, multiplying large numbers. He also provides more technical examples from the cognitive psychological literature. One example involves what he calls "the multi-channels paradigm", according to which any natural scene can be decomposed into, and represented in terms of, four parameters of sinusoidal gratings. For the perceptual psychologist studying form perception within the context of this theory, the game is "to identify formal primitives that adequately describe the visual environment, and to specify algorithms which apply to these primitives to determine complete visual scenes" (Wilson, 1994b, pp. 363-4). Thus, the formal (computational) system in question is not contained within the head of the cognizer.

As a first run at this example, notice that it deals entirely with perception, and there is no reason at all to suppose that anything we might find in it that favors systemic anti-individualism will generalize to the general case of more central cognition involving reasoning and the like. My main problem with it, however, goes straight to the heart of the example and others like it. This example, in my view, illustrates a cognitive (in this case computational) system that is wide (externally individuated) in what we may call the *trivial* sense. Wilson describes the example in terms of what sort of formal system the psychologist is studying. The common formal denominator of this so-called system, which includes internal and external portions, is that all of it is characterized by the psychologist in formal terms. However, the only computation that is going on is going on on the inside. To the extent that the system is computing anything, it is only the internal portion that is doing any computing. It is the internal mechanisms that drive the process. The formally described external environment functions merely as input to the only system that is doing any computing, namely, the nervous system of the animal in question. The "system" that includes internal *and* external portions, is wide in only a trivial sense. And given the means we now possess for distinguishing the psychological from the external (e.g., servomechanisms, intrasystemic stability, etc.), there is little reason for the individualist to feel threatened by this sort of example.

Moreover, even if there were interesting computations being performed by mechanisms external to the individual, that would be no reason to regard these external computations as cognitive; computation by itself offers no reason to think of a system as cognitive. It is a well-worn truth that adding machines are not thinking (at least not merely in virtue of being adding machines). Thus, it is difficult to see why any examples such as this should tempt us to regard processes external to individuals as in some way cognitive.

Wilson offers a second example that involves the way certain animals navigate through environments by means of maps that are continually updated. This involves considerable interaction between organism and environment. But, as we saw in the previous section, and in the discussion of Haugeland earlier in this section, mere intensity of interaction is insufficient to wipe away the distinction between the two interacting systems. I see no reason here to disregard the individuative criteria we have deployed throughout our critical appraisal of systemic anti-individualism.

There are other examples of computational systems that are supposed to be wide in a similar sense. Kaelbling (1993), for example, presents a picture of so-called embedded autonomous agents. Such an agent can be defined formally in terms of a 4-tuple <I,I,R,B>. I is the set

of possible inputs from the world to the agent; I is a function that determines what inputs the agent will get given the states the world is in; R is a reinforcement function; B is the behavior (a function from some stream of inputs to a set of possible actions.

This formal definition of embedded agents is such that two agents can be in the same internal states, yet be formally distinct in virtue of having different I functions (since I functions map states of the world onto inputs). Losonsky (1995, p. 567) makes much of this feature, and indeed makes this astonishing claim about the relation of Kaelbling's model of embedded agents to the Twin Earth thought experiments of Putnam and Burge:

> "What the philosophers tried to achieve by focusing on referential differences, Kaelbling's model captures with her model of an agent. Referential differences, insofar as they themselves are identical to or supervene on causal differences, will be captured by a mapping I. This is a nice example of how independent empirical work can vindicate *a priori* philosophical speculation". (Losonsky, 1995, p. 567).

What is astonishing in this claim is the assertion that Kaelbling's model "vindicates" Burge's original arguments for anti-individualism. This is astonishing for two reasons: First, even individualists who would interpret the thought experiments differently from anti-individualists take for granted that it is at least nomologically possible for two agents to be internally identical, yet receive different inputs. So, if all Kaelbling's model has done is describe in formal terms how two agents can be internally identical and yet receive different inputs, then all Kaelbling's model has done is "vindicate" a step in the argument for anti-individualism that individualists never contested. This is not much vindication in my book.

Second, if the point is meant to be that the agents are *formally* distinct in spite of their internal identity (and hence cognitively distinct as well??), then Losonsky's claim is astonishing because it bears on *systemic* anti-individualism rather than the *individuative* anti-individualism with which Burge is concerned. That is, the claim is now that Kaelbling's model is defined in such a way that formal system includes states in the outside world (much in the same way as Wilson's so-called wide computationalism). Burge, on the other hand, was concerned with how to individuate internal states in virtue of their relations to external states; no such individuative claim is being made by Kaelbling on this interpretation of her work. But if this is what Kaelbling's model comes to, then it is irrelevant to the points Burge intended to make, and therefore can hardly vindicate them.

And third, if this is all Losonsky is claiming on behalf of Kaelbling's model, there is a certain banality, and hence triviality, to his claim; for the claim now asserts only that one can give a formal definition of an agent as essentially embedded in its environment. Who would question that? Not an individualist. Presumably, the claim is meant to challenge Segal's (1991, p. 492) claim that the whole agent minus the environment constitutes the largest integrated system available. But note that Segal's claim is not simply that claim that the whole agent minus the environment is the largest *system* available (for study, for formal characterization, etc.); Segal is well-aware that there are ecological sciences, social sciences, etc., sciences that study and characterize interactions between agents, other agents, and environments. Rather, Segal's claim is a claim about the largest *integrated* system available. Here I wager that Segal is exploiting some of the ideas articulated above in section 1. For example, the cranium and skin serve to preserve the *integrity* of our internal computational system in the face of changing environments. The agent/environment system may admit of formal characterization, but it is not integrated in the way that the agent-internal system is integrated.

We must not forget that virtually any phenomenon can be given a formal characterization. Fires, floods, hurricanes, digestive systems, etc., can all be simulated computationally. Moreover, an agent's interaction with these other types of system can also be characterizes formally. But, of course, just because a phenomenon can be defined computationally does not mean it is cognitive. Once again, we must ask what parts of this so-called wide computational system are cognitive, and what parts are not?

Still more purported examples of systemic anti-individualism can be found in the literature on developmental psychology. Losonsky (1995) has suggested Karmiloff-Smith's (e.g., 1992) work on child development, and Rutkowska's (e.g., 1993) computational infant both contain important evidence for systemic anti-individualism. What sort of evidence is this?

Let us take Karmiloff-Smith's work first. Losonsky cites many different places where Karmiloff-Smith discusses learning that requires complex interactions with the environment. The list includes: The procedure used to solve the Rubick's Cube puzzle requires one to manipulate the Cube and keep track of its changing states; the way children come to understand the relationship between gravity and torque involves moving a block along a balance and feeling the direction of fall until the balance point is located; the solution of problems through the use of "external handles" (such as the use of notational systems like drawing and writing, and distinguishing between

pretending and believing by varying voice, intonation, and body-movement) *extend* the representational structure of the mind; and in learning to tease, children use the adult's external laughter to "change and complete" their representation of the teasing. (See Losonsky, 1995, pp. 361-2.)

Apart from the rather adventurous metaphysical speculations of both Losonsky and Karmiloff-Smith, I fail to see anything here that would constitute evidence for anti-individualism. Losonsky suggests that in the complex interaction between the person and the Rubick's Cube, the Cube itself is part of the problem-solving procedure in virtue of the fact that later interactions with the Cube depend on the way earlier interactions with the Cube have gone. But how does this suggest that the *Cube* is part of the problem-solving procedure in such a way that the cognitive system now includes not just what is going on in the head, but the Cube itself? A more natural way to express what is going on is that *representations* of the Cube are part of the problem-solving procedure. The Cube itself is what the cognitive system, with its internal problem-solving procedure, must interact with in order to solve the puzzle. On this way of looking at the matter, the cognitive system remains within the head, even while it is acknowledged that that cognitive system must interact with its environment in certain well-specified ways in order to arrive a solution to the puzzle. The point that must be emphasized is that there is absolutely *nothing* in the claim that an individual must interact with the environment to accomplish some task that suggests that the cognitive contribution to the solution to the task includes that part of the environment that the individual must interact with. Exactly nothing is lost in this account if it is maintained that the cognitive contribution is found entirely within the individual, and that it is a cognitive system interacting with its environment that leads to the eventual solution to the puzzle. Indeed, described in this way, we can make sense of the fact that it is an *interaction* between two systems. Losonsky, like other anti-individualists, appears to assume that individualists are given to deny that cognitive systems interact with their environments; why else would interactions with the environment be taken by anti-individualists to pose trouble for individualism. But, quite obviously, no such denial is a part of the individualists' thesis. They simply maintain that it is an interaction, not an *intra*-action.

Precisely the same points apply to the other purported bits of anti-individualistic evidence from child development. Children interacting with blocks and scales, the individualist would maintain, simply interact with those external features. In distinguishing between pretending and believing, and in using notational systems, children again interact with the so-called external handles. In learning to tease,

it is not the adult's laughter that completes the child's representation of the teasing incident, but the child's *representation* of the laughing adult. In each one of these examples, no explanatory power at all is lost by conceiving of the cognitive system as contained within the child, and that that system merely interacts with non-cognitive features external to that system. Thus, it is very difficult to see how any of this so-called empirical support for anti-individualism is anything of the sort.

Now consider the putative evidence from Rutkowska's computational infant. Her main point is that child and environment constitute components in a complex feedback loop.

> "...the mind is for action -- internal 'models' are only one component of a system that is designed for acting and developing in the (physical and social) world" (Rutkowska, 1991, p. 98).

Even Marr's 2.5 D sketch is tied to action. In general, the mind affects the environment in ways that, in turn, affect the mind in ways that make certain cognitive developments possible. And all of this can be modeled computationally (hence the title "The Computational Infant").

Note, first, that the very quotation from Rutkowska herself is a bit fishy. What exactly is it that is designed? And what exactly is it that is acting in the physical and social world? And what exactly is it that is developing in the physical and social world? It strikes me that the most natural answer to all of these questions is that it is the child that is doing all of these things. There is, once again, the child *and* her environment; two systems interacting with one another. What is to prevent us from claiming simply that the child's cognitive system has been designed (in Rutkowska's words) to interact with her environment? There is nothing in the idea that two systems interact with each other that even tends to suggest that if one of them is cognitive, the other is as well. All of the evidence we have here is about two types of system interacting with each other. What is remarkable in all of this is that an anti-individualist thinks that there is anything here to which an individualist would object.

It may be that the anti-individualist idea is that if cognitive systems require interaction with an environment to develop, solve problems, etc., then such systems are essentially embedded systems. One simply could not be a Cartesian solipsistic point, devoid of some environment with which to interact. Perhaps there is some contingent sense in which this is true; as a matter of contingent fact, we are the sorts of creatures that require interaction with our environments in order to develop and accomplish what we do. But individualism does not deny this. As far as the individualist is concerned, we *are* embedded creatures; we causally interact with our environments, and we develop

and think in the way that we do as a result of this causal interaction with the environment. The individualist insists only that the properties, facts, and features in virtue of which we are *cognitive* creatures, indeed the cognitive creatures that we are, are properties, facts, and features of us considered independently of our environments. The embedding and causal interaction are, for the individualist, facts about us, but only contingent facts. Had Nature and her laws been different, there is no inconsistency in imagining us to be just as we are.

Perhaps the distinction anti-individualists are missing is between individuation and causation. There is a sense in which the evidence cited by anti-individualists shows that it is in virtue of the environment that we develop and operate the way we do. But this is a *causal* sense of 'in virtue of'. The individuative sense may or may not have anything to do with how we have been caused to be as we are (money, for example, is what it is in virtue of having been minted in certain specific places). But the anti-individualist is not entitled to assume that these causal interactions are also individuative of cognitive creatures. At bottom, we *could* be BIVs, and all these environmental features that the anti-individualist takes to be essential to our cognitive status could be simulated. But if they could be simulated, and we would still be cognitive creatures even if we were BIVs, then it cannot be essential to our cognitive identity that we interact causally with our environments in the way that this evidence suggests we do. Absent some argument, the anti-individualist cannot assume that individuation should respect causation.

4.5 Against Embodiment

The critical remarks have so far dealt only with the attempt to stretch the domain of psychology to include the environment. Part of the case for systemic anti-individualism, however, is directed at undermining the distinction between mind and body; given the high bandwidth interface between body and world, this is tantamount to undermining the distinction between mind and world. Haugeland in particular aims to show not just that the mind is intimately related to the brain, but that the mental, insofar as it is involved with the brain, is just as intimately involved with the body. Borrowing points broached initially by Dreyfus, Haugeland tries to expose the comminglingness of brain and body essentially by appealing to the messiness of the interface between brain and motor neurons; to draw a sharp distinction here would be like disentangling a bowl of spaghetti (not Haugeland's analogy). Neural projections from the brain reach out to a bewildering array of motor neurons in our muscles; and these are in constant informational flux as a result of their high bandwidth interface with the ever-changing environment. There is simply no way that "simple instructions" or

"well-defined, repeatable messages" could govern the richly textured interplay between mind, body and world. In Haugeland's eyes, we can no longer sustain a picture of "...two relatively independent separable components -- a rational mind and a physical body, meeting at an interface -- but rather a closely-knit unity" (Haugeland 1993, 22).

But work on the neural organization of motor control belies Haugeland's worries. (Gallistel 1980, esp. 275-280) has reported substantial evidence of a hierarchical organization in motor control. This is significant because it has the consequence that an initial motor command need not specify the details of its implementation in motor behavior (Ghez, 1985; Ghez and Fahn, 1985). The reason for this is just the same reason an army General need not stipulate how each troop is to behave when he hands down the order to assault. The General tells the Colonels, the Colonels pass the word on down the chain of command; the specific details of the implementation for the command are left to the individual troops. The same is true in the neural hierarchy that governs motor behavior. Through a series of (sometimes nested) servomechanisms, feedback information from the environment is accommodated by lower-level control centers (analogous to Majors and Captains in the military chain of command).[18] The top level of the neural hierarchy sets out the instruction; the lower levels fight it out with the environment to determine specific details of implementation. The same organizational structure, moreover, can be extended to cover rather more sophisticated behaviors of the type normally taken to issue from beliefs and desires (Butler, 1992).

But now we can see how the spaghetti-like interface between brain and body might admit of a structured over-lay. Haugeland is skeptical that any set of instructions "from the top", as it were, can adequately for the behavior that results from the high bandwidth interface between brain and body. But our recognition of the motor control hierarchy allows us to make sense of central control in a way that Haugeland cannot. The presence of a high bandwidth interface between body and world has no relevance to the intelligibility of well-defined, repeatable instructions encoded in the brain. Servomechanisms in a hierarchically structured control system insulate the brain, and the mind, from the neuromuscular spaghetti, and hence from the outside world. A hierarchy of servomechanisms, in other words, creates a *narrow* bandwidth interface between brain and muscles. In light of this, we can appreciate why it is that a high bandwidth interface between body and world does not draw the mind out into the world.[19]

Similar points can be used to defuse a related trend now emerging in the analysis of situated cognition and dynamical systems approaches to cognitive sciences (see, e.g., Clark, forthcoming). The suggestion here is that organism and environment are often "coupled" in an active causal

interchange; this interchange may be low bandwidth, in Haugeland's terminology, but it is so fast and fluid as to effectively blur the distinction between organism and environment. Swimming, for example, is best viewed as a coupled system consisting of swimmer and water. The rapid rate at which the swimmer gets and responds to information about the fluidly changing environment obviates any attempt to understand the behavior as the product of a coherent internal plan. It is as much the product of the environment as it is the swimmer.

As it turns out, however, the rapid causal/responsive interchange is confined almost exclusively to the lower reaches of the servosystemic structures insulating the swimmer's mind from her environment. The swimmer forms some sort of an action plan that is carried out with localized adjustments to her fluidly changing environment. She need not attend to the adjustments in any conscious way (the signals controlling the adjustments may never reach higher than the spine or brain stem). Thus, while the swimmer may well be coupled with her environment, that fact is not necessarily of paramount psychological interest.

We can see now why a psychological explanation of non-autonomously described behavior should not involve the peculiarities of an individual's musculature and the near-indeterminacies of the environment. An explanation must provide an understanding, and to do so, it must abstract from details. High bandwidth interfaces and rapid responses simply contain too many details to be intelligible; they *prevent* us from finding what we are looking for in a psychological explanation. An individual's beliefs and desires, even her algorithms and computations (see below), are protected from her body and world by a hierarchical nexus of servomechanisms. Haugeland is quite right that high bandwidth interfaces characterize many of the divisions we have heretofore endorsed uncritically, but our own assessment affords us those same divisions upon further review.

4.5 Anti-individualism and Naturalism

None of the views we have looked at so far succeed in showing that if cognitive systems are to be seen as part of the natural order, they must be taken to *involve* the environment rather than to be separated from it. As a final case, consider the remarks of McClammrock (1995), who claims that the semantic and subjective character of the mental can be understood only when cognition is taken to be world-involving and embodied. McClammrock (1995, e.g., p. 166) suggests, for instance, that understanding how states acquire their semantic or representational properties, their content, requires us to pay attention to "the implementation of the embodiment" of the states that have that content. He suggests several features of an embodied and embedded

brain that might help fix the content of the brain's representations. Take, for example, the fact that there are usually several different causal pathways leading from external object to a representation of that object. This is taken by McClammrock, and before him Dretske (1981), to show that it is the external object that is the object of representation rather than any more proximal stimuli. But, as Dennett (1987, p. 311ff), among others, has shown, this multiple pathways "solution" does not solve the problem of intentionality any more than mere stipulation solves the problem. We are left with the question, Why is it that this single external object is the object of representation rather than a disjunction of proximal stimulations that correlates perfectly with the presence of the external object? The multiple pathways suggestion identifies the problem without solving it.[20]

As for subjectivity, there is little reason to think that an account of it will be essentially world-involving. McClammrock's (1995, ch. 11) case for the situated nature of subjectivity involves simply the reference-based anti-individualism considered at length in the Part I of this book. The subjective nature of the mental might be externally individuated if our concepts, those essentially subjective states, are externally individuated. I have nothing new to add to my critical remarks of this sort of anti-individualism.

I am defending the sanctity of the cognitive in terms of features that cannot be accounted for in terms of something *else*, something environmental or relational. In so doing, however, I am not claiming that cognition and thought are outside the natural order. There is, in other words, nothing non-naturalistic about the individualism inherent in Cartesianism. Functional properties are not reducible to physical properties, and physical properties are not reducible to anything else; no one is tempted for these reasons to claim that properties of these sorts are non-naturalistic. Why, then, should the fact that mental properties (as opposed to mental substances) are not reducible to something else entail that they are non-naturalistic? The repeated insinuation that some sort of systemic anti-individualism is demanded by naturalism seems to me to be wholly without merit. Indeed, if the remarks of the previous chapter are on target, it is anti-individualism that faces trouble from the naturalistic flank.

From one point of view, McClammrock's project, like Haugeland's, C&C's, Wilson's, and those of other like-minded anti-individualists, can be seen as attempting to rob the cognitive of its distinctive characteristics by maintaining that, to varying degrees, there is nothing special or distinctive about cognition; it is just one more physical or functional processes. On the other hand, we can see the systemic anti-individualist as embracing the antithesis of eliminative materialism. Rather than eschew a privileged conception of mind as a

misguided postulate of an outmoded theory, they endeavor to find it everywhere -- in our muscles, in a beach, and even on the road to San Jose. In fact, they would like to find intelligence in our tools, our institutions, and our culture. We might call this a *ubiquitous* materialism. Intelligent cooperative activity is everywhere around us. For example,

> "...the structure of an institution is implemented in the high bandwidth intelligent interactions among individuals, as well as between individuals and their paraphernalia. Furthermore, the expertise of those individuals could not be what it is apart from their participation in that structure. Consequently, the intelligence of each is itself intelligible only in terms of their higher unity" (Haugeland 1993, 32).

Intelligence is here portrayed as a sort of emergent property, the bases of which, in part, are the intelligences of individuals. But this cannot be right; when collections of things interact to give rise to what we might timidly call an "emergent" property, it is rarely if ever the case that the property that emerges is anything like the properties that it emerges from. Individual water molecules are not wet, but a big-enough collection of them is; individual neurons are not smart, but a big-enough collection of them is. Surely the interaction of individual intelligences, if it gives rise to any further property, does not give rise to more intelligence; such a supposition, if nothing else, has no historical precedent.

In this quotation we see a glimpse of why the systemic anti-individualists' project can seem so wrong-headed: It tries to turn psychology into a kind of anthropology or sociology or ecology; and it just won't fit. There already are sciences whose topic of inquiry is the interpersonal and environmental. More importantly, none of the concerns that systemic anti-individualists raise diminish (in substance or significance) the problems inside the head that psychology traditionally takes on. There is no room for an expanded psychology, no motivation for it, and no need for it. No one will deny that appealing to environmental information can be of great help in unlocking the secrets of the mind, but this, again, is a *methodological* point, not a *metaphysical* one.

This point is often, *very* often, overlooked. McClammrock (1995, p. 192), for example, identifies his view with the "explicit rejection of the Cartesian view" of a mind that is separated from the world in which it (contingently) exists. A similar sentiment is expressed throughout the work of C&C and Wilson. One can, however, recover all of the methodological insights (of, e.g., Gibsonian psychology, Marr's

theory of vision, Ballard's animate vision paradigm, etc.) without making the further metaphysical claims about the nature of minds or cognition. That is, looking at the way organisms interact with their environments and each other will surely help investigators determine what it is that is going on inside the heads of those organisms. To draw the further metaphysical point, however, that cognition literally extends to these external relations or should at least be individuated by appeal to such relations, one would have to produce a rather sophisticated theory of how explanatory projects are to be interpreted in arriving at claims about how cognitive systems are individuated. Such an analysis might be just around the corner. On its own, however, the metaphysical claim outreaches the explanatory and methodological support.

5. Concluding Remarks

I have given my reasons for thinking that thought and cognition are internal affairs. In contrast to the picture of the cognitive that I am advocating, however, Tuomela seems to adopt a very liberal view, and even assimilates psychology to the social sciences:

> "...[Individualism] does not seem to spare much of social psychology and, more generally, of social sciences. ...[C]onsider social relations such as two persons' loving each other. Our ordinary descriptions of such activities and relationships are clearly noninternal. For instance, if John loves Mary, she must exist; if John talks to Mary, she must exist, and so on. It seems difficult to do very much with only internally described psychological states and activities." (Tuomela, 1989, p. 43.)

Tuomela is of course correct in thinking that an individualist psychology will not be able to explain such social facts. Nor, of course, is it supposed to; I am not claiming anything about social psychology or other branches of psychology whose concern reaches out beyond the cognitive. There is no science that explains everything about a given event or set of events. Astronomy explains certain facts about planetary motion, but there are purely physical explanations of those motions, too, explanations that account for those aspects of the motions not relevant to an astronomical account. Economics explains certain aspects of monetary transfer, but there are other aspects left unexplained economically; sociology, psychology, physics, etc., are brought in to explain these further aspects. But the fact that these other scientific factors account for certain aspects of monetary behavior does not make them all economic factors. For exactly the same reasons, the fact that noninternal factors

are brought in to account for social relations, and even for more personal-level behavior, does not make them psychological factors.

But what, then, can we use to individuate scientific domains narrowly enough to draw the divisions between sciences that I am advocating? It is incumbent upon me to offer some criterion by means of which to justify my preferred level of grain. I do not think that the question is as deep as it might first appear. The answer has to be given in terms of laws and lawlike generalizations. Scientific domains are individuated by laws and generalizations linked by a common vocabulary. These generalizations must enjoy a certain autonomy with respect to generalizations at different levels, but it is a mistake to think that they must always be featured alone in explanations. The scientific integrity of economics depends on the existence of such laws, even if the economic explanations in which they normally appear are also committed to contextual assumptions from other sciences. The situation seems to be no different with an individualistic psychology. The fact that individualistic psychology often, even usually, conspires with other sciences (by making various contextual assumptions) in no way entails that those other scientific factors are psychological factors.

The debate over individualism in the cognitive sciences is, then, no mere terminological squabble. There are substantive claims at stake about whether individualism can yield projectible generalizations. I think such a yield is already in place. For example, once we realize that the proper explananda for an individualist folk psychology are other psychological phenomena, we recover all the psychological laws and generalizations we would expect a scientific psychology, or even folk psychology, to deliver. Beliefs or belief-like states conspire with desires or desire-like states to cause intentions to satisfy the desire in some way or another. This seems to me to be almost obvious. The only trick is to notice that these individualistic laws are often invoked in explanations that make substantial internal and external contextual assumptions. It is only the failure to analyze standard behavioral explanations properly that fosters the illusion that external factors are somehow psychological.

The fact that no good argument for anti-individualism has emerged places the view in a uniquely disadvantageous position. It is not in dispute that events internal to an individual are involved in cognitive activity. Only the very brave would dispute that there are some cognitive states and processes that supervene solely on the internal (even if there are some who would maintain that some states and processes do not supervene on the internal). Taking the internal to be cognitive, therefore, requires no further argument. I have even suggested what I take to be distinctive about the internal in virtue of which it is cognitive. What is controversial, however, is the claim that there are

cognitive states and processes that are individuated by appeal to, or even extend to include, an individual's historical or environmental context. If it is to be rational to accept this view, then a compelling argument is required. As far as we have seen, though, there are no such arguments. (Indeed, when it comes to propositional attitudes and causal powers, we have seen that there are in fact good reasons to reject anti-individualism outright.) This leaves the anti-individualist, unlike the individualist, with no firm footing. Thus, in neutralizing anti-individualist arguments, we have effectively sealed its fate. Unless and until we are given some reason or another to take cognitive states to be individuated relationally, or to extend into the environment, we must regard our thoughts and cognitive lives as internal affairs.

NOTES

INTRODUCTION

1. In point of fact, I think there is a jump from the claim that our beliefs could be entirely mistaken to the conclusion that cognition is an internal affair. Nevertheless, Descartes' thought has seemed to suggest a form of internalism to subsequent generations of philosophers. Rather than worry about Cartesian scholarship and what Descartes might have had in mind regarding the individuation of mental states, I shall simply pass on to the claims that constitute his purported legacy.
2. For the most part, I will use the term "externalism" because it is shorter. In certain areas of this book (e.g., Chapter 4), however, I will use "anti-individualism", rather than "externalism" and "individualism" rather than "internalism", in order to maintain contact with the works I will be discussing. Some (e.g., Egan, 1992) distinguish between externalism and anti-individualism; externalism, according to her, is a view about how the contents of mental states are individuated, while anti-individualism is a view about how mental states are individuated (where this may not be by contents). There is also a tradition that takes externalism to be about linguistic content, while anti-individualism is about mental content (or mental states *simpliciter*). My concern throughout will be with the individuation of mental states; for the most part, I will be concerned with their individuation by appeal to their contents, though I will also be concerned with other methods of individuation. When it is important to distinguish externalism in this sense from externalism in epistemology, I shall use the expression "psycho-semantic externalism".
3. Internalists are not the only ones who might appeal to a supervenience relation in understanding the ontological status of the mental; externalists such as Heil (1992, ch. 3) claim that externally individuated propositional attitudes supervene on the environment (or one's history of relations to the environment). The ultimate plausibility of Heil's view, I think, depends on whether mental causation can be understood in externalist terms; this is the topic of chapter 5 below.
4. It was brought to my attention by Eric Saidel that barnacles will cling to a wharf long after they have died. It is not my intention to cling to my internalist intuitions if they are shown to be doomed. The trouble with similes is that they extend only so far.

CHAPTER 1

1. Externalism can also be defended by abstract considerations coming from the theory of reference, and by a direct argument for a particular form of externalist view like Dretske's information-theoretic approach to the semantics of mental states. Burge (1989b, 1994) expresses sympathy with considerations from the theory of reference, and seems implicitly to adopt an informational view in the tradition of Dretske. I will have something to say about each of these, though my main interest will be in "twin" considerations that are featured prominently in Burge's work.
2. I am, for the moment, suppressing Putnam's reasons for favoring an indexical interpretation of the semantics for natural kind terms. For critical discussion of that view, see Burge (1982, pp. 104-6).
3. There are two distinctions that many have found useful in characterizing the nature of the supervenience relation. The first, global vs. local, concerns whether supervenience characterizes the relation between the entire set of physical and mental properties, or whether that relation can apply to individual properties or an individual's (mental and physical) properties. The second relevant distinction, strong vs. weak, concerns whether the supervenience relation is logical (holds in all possible worlds) or merely nomological (a function of the laws in this particular world). For my part, I believe that mental states supervene locally (on

states of the individual's brain), and I am inclined to think (but do not want to commit to the idea) that they do so strongly (so that there is no possible world in which I am in this brain state but thinking a different thought). There are complex issues here that I want to resist getting into for fear that I will be distracted from my main task in this essay, which is to consider whether violation of either strong or weak local supervenience is tenable. For superb and detailed discussions of the supervenience relation, see Kim (1982, 1984, 1987), Heil (1992, ch. 3), Horgan (1993), and Chalmers (in press, ch. 2).

4. Here is as good a place as any to introduce the usual caveats about the ontological status of concepts and contents. The ontological status of supervening properties is of great metaphysical interest. I want to remain, however, as neutral as possible on this issue. Though I approach the issue from a materialistic and even naturalistic point of view, that conviction functions *at most* as a motivation to find support for a supervenience claim that binds mental activity to physical activity in a way that is not ultimately mysterious. I intend for my arguments against externalism to go through even if my materialistic and naturalistic intuitions cannot be massaged as nicely as I would like.

5. Burge (1979, pp. 82-85) argues that it is not necessary to appeal to incomplete or incorrect understanding of the literal meanings of terms used obliquely in attitude attributions; even concepts associated with fully and correctly understood terms depend on the social environment. The argument consists in running the thought experiment in reverse, so that the misunderstandings and partial understandings occur in the counterfactual case rather than in the actual case. I believe that the plausibility of the reverse thought experiment is parasitic on the plausibility of the original thought experiment, in which case some partial or incorrect understanding is required. However, as we will see in a moment, Burge (1986c) does offer a thought experiment and attending argument to his externalistic conclusion that seems not to depend on ignorance of the literal meaning of obliquely occurring terms.

6. Bach (1982, pp. 129-131; 1987, ch. 10; 1988, p. 95) makes this point quite effectively in terms of specific arguments about the semantics of belief reports. The general point Bach makes is that in the ascribing of a belief content to an individual, the ascriber sometimes uses the terms at oblique occurrence in the ascription to express how she, the ascriber, thinks or conceives of the referent of the believer's thought. Jacob (1990) seems to adopt a similar position.

7. In Butler (1995a, b) and in Butler and Shogenji (in preparation) I defend at great length the compositionality of mental content.

8. The fact that individuals have difficulty recalling the analyses of their concepts (e.g., providing definitions on demand) can be explained by the fact that these analyses are progressive, and therefore extensive. Whatever cognitive resources there are for such self-analysis, they can surely be swamped by the sheer volume of data with which they would have to contend. Thus, limitations of introspective knowledge do not count against the view that complex concepts are compositional constructions out of basic concepts. Note also that these limitations on introspection are independent of the limited claims of self-knowledge discussed in chapters 2 and 3.

CHAPTER 2

1. See, e.g., Burge (1986, 1988a, 1988b); Brueckner (1987, 1990, 1993, 1994); Heil (1988, 1992 ch. 5); Boghossian (1989); Falvey and Owens (1994).

2. See also, e.g., Bilgrami (1992), Bonjour (1991), Greenwood (1991), who agree that if externalism implies that we lack self-knowledge, it faces rejection by *reductio*.

3. Millikan (1994) presents a compelling way to understand Russell's principle, while avoiding the apparently incoherent interpretation on which acquaintance with the object of one's thought requires one somehow to step outside oneself and discern a match between thought and object. Her notion of "coidentifying" requires

only that one have the ability to match (to varying degrees of reliability) two thoughts of the same object as somehow being two thoughts of the same type. Doing this infallibly, of course, would require one to have something like the impossible ability Russell's principle seems to require. But doing it at all, and doing it well, may be accomplished without that requirement. It is simply a matter of knowing when two thoughts *purport* to be of the same object.

4. Following others in the literature, I will use this generic term "thought" to refer to propositional attitudes without making a commitment to any particular attitude.

5. Bernecker (1996), however, presents an interesting argument for the claim that externalism is inconsistent with knowledge of the attitude component of thoughts.

6. Some externalists may not welcome this concession. Burge (1988b, p. 658) claims that "[n]o errors at all are possible in strict cogito judgments...." F&O (1994, p. 118), too, ask "How can it be that the subject is always right about the contents of her beliefs...?" I will review the plausibility of these claims in sec. 2.5 below.

7. I take it that it is clear that the rejoinder Warfield (1992, p. 235-6) anticipates, which concerns a strong version of *a priori* self-knowledge that is not a species of *introspective* self-knowledge, is irrelevant to the present rejoinder, which *does* concern introspective self-knowledge.

8. Chase Wrenn has remarked (in conversation) that it is not clear that Ludlow's examples involve *slow* switching, where one has a chance to develop a new concept in the new context. Citing a view of Evans (1982) about the development of new concepts on an externalist account of meaning, it may be that the participants in Ludlow's stories lose their old concepts without thereby forming new ones; they would be in a period in which they have *no* concept, neither old nor new. As Wrenn points out, however, even if the stories are to be interpreted in this way, that is no help to the externalist. For on this interpretation, the individuals take themselves to know what they are thinking, but in fact are not have thoughts with any content at all (since the so-called thoughts involve what can only described as contentless concepts). In what sense, then, does one know what one is thinking when one could be mistaken in thinking that one is having thoughts at all? For surely one cannot be said to know the contents of one's thoughts if one is not able to tell whether the purported content is content at all.

9. See, e.g., Heil (1992, ch. 6), Stalnaker (1991, esp. pp. 143-4), F&O (1994), and Forbes (1995).

10. Heil (1992, p. 179) talks, in his discussion of externalism and self-knowledge, about entertaining a thought "self-consciously", and avows that this position is "close to Burge's". (This may be, for Heil, somewhat of an advance beyond his own 1988 position.) In a similar vein, Stalnaker (1991, p. 144) points out that "...the registering of a registering of a blue cube is itself a registering of information about blue cubes." Also, it is likely that F&O (1994, p. 118) have in mind the same idea since they too maintain that self-knowledge displays an immunity to error, and then use this to help motivate their endorsement of the second component of the standard strategy (as it is described in the following paragraph).

Dretske (1994, 1995, ch. 2) offers an information-theoretic version of at least this first component of the standard strategy (and has in conversation embraced the second component). On the information-theoretic account of representation, an internal state, Mp, represents an object, O, if and only if it is the function of M, when it is in state p, to carry information about (or indicate) O. M has the function of carrying information about O if, when it is in proper working order (where this, presumably, is determined by its adaptive/biological function), it is in state p when O's are present. For a representational system, S, to know what it is representing, then, it must have access to information about how the world would be when M is functioning properly. But that information is available simply in virtue of the fact that Mp represents O (since it would not represent O if O was not present when M is properly functioning and in state p). In terms that recall Stalnaker's claim, the representation of a representation of X represents X. This does not entail that S does *in fact* represent how it itself represents the world, but only that it is possible for S to do so, insofar as the resources necessary for doing so are available. So, as

with the less technical versions of the standard strategy, if one meets the conditions for tokening a representation with the content that O is present, then one is able *ipso facto* to meet the conditions for having a second-order representation involving that content. And, since there would be no room for error, there is reason to take this second-order representation to be a case of knowledge.

11. F&O use the phrase "knowledge of comparative content". It has been suggested to me that it would be better to use the phrase used in the text since it captures the idea at issue better than F&O's own phrase.

12. Note that on the surface, the move from (4) to (5) might seem to depend not on Closure, but on a similar principle that we might call Negative Closure:

$$[\neg Kp \ \& \ (Kp \rightarrow Kq)] \rightarrow \neg Kq$$

(This appearance is due to the fact that both (4) and (5) involve claims about what S does not know.) This principle, of course, is false, so if that is indeed the reasoning, even Brueckner's version of the argument is bad, and hence uncharitable, perhaps suggesting that no charitable version of the argument is forthcoming. As a matter of fact, however, we will see in detail below there is an obvious way to get from (4) to (5) using a *reductio* that embodies the form of argument dictated by Closure. Therefore, the charge of uncharitability does not stick against Brueckner.

13. Ludwig (1992) endorses an argument essentially like this (cf. pp. 234-5).

14. I note in passing that F&O's treatment of RA' (pp. 114-118) occurs in their article *before* they discuss Brueckner's content skeptic (pp. 118-123). That alone should suggest that they did not intend to canvass the content skeptic's reasoning along the lines of RA'.

15. F&O may not have anticipated Warfield's resistance to the relevance of twin thoughts, but we have seen in section 2.2 that Warfield's resistance is in any case misguided. Thus, F&O's concession is in no way problematic.

16. The following line of argument is in sympathy with the one given by Boghossian (1989, pp. 20-23, esp. p. 23). The argument presented here, however, proceeds to the same conclusion a little more slowly. My hope is, in part, to remove any reservations one might have about Boghossian's argument.

17. That is, such discriminations must be possible in any case that is not a case of basic self-knowledge. In the basic cases, the impossibility of error, and hence the status of the second-order belief as knowledge, is guaranteed by the fact that the first-order content is contained in the second-order content. Since this guarantee is independent of the ability to rule out alternatives, even just relevant alternatives, it survives the inability to rule out such alternatives.

18. Burge (1988b, p. 658, p. 663, fn. 11) claims that the standard strategy generalizes comfortably from the pure forms of basic self-knowledge, or at least suggests a comfortable generalization, to cases that do not involve infallibility. We are seeing now, however, that this appears not to be the case. The standard strategy is of no help in accounting for anything but the basic cases, and its failure in the non-basic cases ricochets back to impugn the account it provides in even the basic cases.

19. This is one way to understand Boghossian's claim that externalism treats self-knowledge as if it is cognitively insubstantial (see, e.g., Boghossian, 1989, pp. 17-23). Consider self-verifying judgments such as "I am here now", or "I am jealous". Anyone who makes such judgments knows them to be true without the help of any empirical evidence. Making such judgments knowledgeably, then, involves no "cognitive achievement". Cognitively substantial knowledge, on the other hand, involves gathering evidence, or at least putting forth effort to acquire information, and sometimes (in cases not involving knowledge of content) forming a judgment based on that evidence or information. Such knowledge involves cognitive achievement. The standard strategy maintains that basic self-knowledge, insofar as it is infallible in this way, involves no cognitive achievement, and is therefore cognitively insubstantial. However, it is not generally plausible to treat knowledge of content as being cognitively insubstantial. Even basic self-

knowledge can be more or less thorough; children, for example, seem in general to be somewhat less proficient at culling information from introspection. The amount and reliability of introspective information seems also to be a function of the extent to which one pays attention to one's thoughts:

> "...self-knowledge is both fallible and incomplete. ...[i]t is only if we understand self-knowledge to be a cognitive achievement that we have any prospect of explaining its admitted shortcomings." (Boghossian, 1989, p. 19.)

This is not to say that thoughts about content are not direct, or that they are not more authoritative than empirical thoughts. It is just that Cartesian ruminations about the pristine authority of self-knowledge are exaggerated in the extreme; even basic self-knowledge is cognitively substantial, and any account of it must allow for this.

20. In suggesting that Mates thinks antecedently that his two thoughts have the same content, I do not intend to be appealing to a notion of content that outstrips the language that might be used to express it. We may suppose that (i) is the correct linguistic expression of Mates' thought contents at t_1 and t_2. Mates' views on language affect the beliefs he has about any *further* expressions that might be used to characterize his thought linguistically; specifically, he thinks that the thought he has at t_2 is expressed by (i), but he *also* thinks it can be expressed by (ii). He thinks this because he thinks his thoughts at t_1 and t_2, namely the thought expressed by (i), are the same.

I thank Carolyn Morillo for helpful discussion on this point.

21. It is worth introducing a caveat here: I am assuming that Mates and Church share their understanding of the meaning of "sameness". This assumption may be false, and I address that possibility below.

22. Recall that we have adapted F&O's own argument so that it avoids the errors pointed out above, so we are here just speculating about how they would describe the case. Their claim that Mates knows what thoughts (i) and (ii) express does not guarantee that if Mates has one or both of those thoughts he will deploy (i) and (ii) appropriately in expressing them.

23. It is not difficult to see that the ability to detect differences in the determinants of contents is not equivalent to knowing what determines contents to be as they are. Just imagine a syntactic engine that detects syntactic differences that correspond to semantic differences. Or, more simply, imagine a thermostat that detects air temperature by means of the amount of curvature in a bimetallic strip. In neither case is detection sufficient for knowledge.

24. Boghossian raises (p. 15) Humean worries about the directness of our knowledge of causal relations. I think such worries are not germane in this context since it is not, in the first instance, the *causal* nature of the relations that we must have direct access to. Rather, it is the mere fact that one's thoughts or concepts are related to others, and there is no reason why we could not introspect such relations.

25. This flies in the face of the notion of so-called "found art"; it is also antithetical to deconstructionism. This is fine with me (though, perhaps, not with others), as I think both ideas have deeply confused and implausible commitments about the nature of representation.

CHAPTER 3

1. See especially Falvey and Owens (1994), Brueckner (1994), and Davies (1994). There has been quite a lot of recent philosophical activity on this and related questions, beginning with Hilary Putnam (1981, ch. 1). See also, e.g.,: Burge (1986, 1988a, 1988b); Boghossian (1989); Brueckner (1986, 1990, 1992a, 1992b, 1992c); Casati and Dokic (1991); Collier (1990); David (1991); Dell'Utri (1990); Gallois (1992); Heil (1987, 1988, 1992); Ludwig (1992); McKinsey (1991); Smith (1984);

Stephens and Russow (1985); Tymoczko (1989); Davies (1995); Warfield (forthcoming).

2. My interest is not in Putnam scholarship; the gloss on the argument that follows, which I will be considering in this paper, may differ from the one Putnam actually intended to give (indeed, see Davies, 1995, for a compelling defense of the claim that Putnam's own argument was intended to have only limited scope). I will be concerned with an "extended" interpretation of Putnam's argument, on which the argument is taken to have more general anti-skeptical implications than may Putnam anticipated, because I believe that the literature that Putnam's own argument inspired has been itself concerned with the extended interpretation.

3. Or that there are "tree" users (cf. Burge, 1982), or that there are things corresponding to the concepts that figure in the construction of the concept of tree (this last qualification is needed to handle terms like "unicorn").

4. I should admit explicitly here that I am relying on a modal interpretation of skepticism, on which it claims not just that we happen not to know anything about the external world *per accidens*, but that we are systematically debarred from such knowledge due to a unbreachable justificatory gap. Such an interpretation should be uncontroversial; arguments on behalf of skepticism uniformly purport to raise problems *in principle* with our ability to justify claims about the external world.

5. This presentation of F&O's line of argument involves two bits of reconstruction. First, in quoting F&O, Brueckner writes (p. 337) that "the truth condition in question concerns 'the *external* states of an organism that normally cause it make a given judgment,'...". The italicized word in the quotation does not appear in F&O's text, which instead contains the word "internal". I can only surmise that Brueckner is right in replacing the word; immediately above, F&O interpret the truth conditions of the BIV's utterances as the normal cause of the sense impression, which are surely external to the BIV, not internal. Second, F&O (p. 129) describe electronic configurations Q, the truth condition of the BIV's utterance of 'I am a BIV' as "the normal causes of a BIV's tokening of the sentence 'I am a BIV'." I am supposing that Q causes the tokening of that sentence by first causing the sense impression that is stipulated to remain constant across BIVs and non-BIVs. This interpretation is supported by F&O's remarks on p. 130, mentioned just above in the first reconstruction, that the truth condition of the BIV's utterance is the normal cause of the sense impression. I am being explicit in noting and justifying these reconstructions here because they will play a role in the proper assessment of Brueckner's critical remarks of F&O's analysis.

6. This is not quite the argument that F&O consider; the original premise (A) began: "I know that...". This preface to the original (A) does no work and should therefore be eliminated. A similar adjustment has been made to (C).

7. It is ironic that Brueckner (1994) has attempted to refute the QB strategy since in his (1992b) he endorses what is substantially the QB strategists' main point in a closely related context. McKinsey (1991), like QB strategists, is dealing with the possibility of a *reductio* of externalism based on implausibly strong anti-skeptical implications. McKinsey's argument is based on the claim that the following three propositions are inconsistent:

(1) Oscar knows *a priori* that he is thinking that water is wet.
(2) The proposition that Oscar is thinking that water is wet necessarily depends upon E (a proposition about the external world).
(3) The proposition E cannot be known *a priori*, but only by empirical investigation.

A key problem with McKinsey's argument concerns the nature of the dependence expressed in (2), and whether it can be known *a priori*. Brueckner (1992b, pp. 114-5) argues that the dependence cannot be known *a priori*. To deny this, one would be required to show without recourse to *a posteriori* knowledge that certain thought contents display a "counterfactual dependence" on certain external facts;

that is, it would be incumbent upon us to show *a priori* that had one's environment been sufficiently different, so would the contents of one's thoughts. Thus, in essence, we would have to establish externalism without making any empirical assumptions. But this, Brueckner argues, cannot be done. We must at some point *assume* something about what the actual world is like, e.g., that the world contains H_2O rather than XYZ, so that we will have a touchstone relative to which we can determine what a variation in external context would be like; since such an assumption would be empirical, it could not be knowable *a priori*.

Brueckner's response here, of course, makes the same point as the QB strategy insofar as it purports to block the anti-skeptical implication of McKinsey's argument by pointing to the *a posteriori* and (hence) question-begging nature of the externalist premise. This point, in my view, obviates an attempt by Brown (1995) to defend McKinsey's argument. If externalism is assumed, then the argument can go through; but externalism cannot be assumed without begging the question against the skeptic.

8. Brueckner (1994, p. 341) wrongly asserts that the remarks about brains constitute an independent argument in support of premise (B) in *S'*. In point of fact, it is merely a continuation of the support that began with a consideration of the case of water and the Twin Earth story; the Twin Earth story and the remarks about water, on their own, have nothing whatever to do with (B), which is about reference to brains, not water. Therefore, the discussion of brains is needed to complete the support for (B) that began with the consideration of the Twin Earth story and water.

9. Brueckner treats this rejoinder as a mere possibility one "might" explore to try to get around his criticisms. In point of fact, I think it obviates his criticisms. I do not know why Brueckner came to think that *a little bit* of assumed empirical knowledge would be all right in the context of an argument against skepticism. I think it can be shown that any argument that applies to the assumption that water is H_2O applies to the assumption that water is a liquid; *mutatis mutandis* for the corresponding assumptions about brains.

Moreover, relating to points to be considered below, the consequence that I can know that liquids or body-parts exist in my environment solely on the basis of introspection is every bit as unattractive as the consequence that I can know by introspection alone that water is H_2O or that I am not a brain in a vat.

10. It has been suggested to me that one need not assume an empirical claim such as the claim that water is H_2O in supporting externalism; one need only imagine two hypothetical language communities, and that *if* one has grown up using "water" to refer to H_2O, then one cannot also use it to refer to XYZ. But I think F&O are right to make the charge of question-begging against Putnamian anti-skeptics: The two-hypothetical-communities version would not be not sufficient to establish the interpretation of the Twin Earth thought experiments on which externalism depends. The externalist interpretation depends on comparing the Twin cases to semantic intuitions that are grounded in actual practice. Without actual practice, and the empirical claims that it involves, there would be no reason to prefer the interpretation that favors externalism over an interpretation that does not. Thus, while the Twin-Earth-type arguments can indeed be restated in hypothetical terms, there is no reason to think that they can be compelling outside of their original statement in terms of actual practice.

11. Because I have received only an electronic copy of the manuscript on which Davies based his presentation, I cannot provide page references. I apologize for any inconvenience this might cause.

12. For reasons that become apparent below, I suspect that, in the antecedent, Davies should have said, and probably meant to say, that the *belief* that B is a precondition, *not* that the truth of B is a precondition.

13. Once again, I suspect that Davies should have said in the antecedent that believing that B is a precondition, not that the truth of B is a precondition. Nothing in the criticism of Davies' argument that I will later develop turns on this point.

14. It is worth pointing out that, contrary to the strategy adopted by Brueckner (1992a) and David (1991), the anti-skeptical argument can be refuted without first claiming that externalism is inconsistent with *a priori* self-knowledge, in which case the argument would be unsound. (Thus, Warfield's, forthcoming, claim to the contrary is false.)
15. Warfield (forthcoming), for example, goes so far as to say (in fn. 2 and fn. 4) that externalism is so plausible as to be beyond question. I find this assertion to be incredible, especially given that Warfield offers no support for it. Nevertheless, my main disagreement is not with Putnamians like Warfield, but with the QB strategists, who are explicit in acknowledging that externalism is vulnerable to a *reductio* if its epistemological implications are too robust.
16. The fact that Putnamian anti-skeptical arguments fail in a way that leaves the anti-skeptical implications of externalism intact prevents one sort of response to the apparent conflict between externalism and skepticism: Some (e.g., Warfield, forthcoming) suggest that the anti-skeptical implications of externalism threaten skepticism, not externalism; most agree that skepticism is false, and Putnam's argument shows that agreement to be well founded. Given the QB strategy and my response to it, however, we are left with the logical tension between externalism and skepticism, yet with no possibility of a compelling argument from externalism to the rejection of skepticism.
17. More generally, one may not be able to *show* that skepticism is false based on the known truth of externalism (and home truths about self-knowledge); it is just that one could not come to believe that externalism is true without *presuming* that skepticism is false. The truth of externalism conflicts with the truth of skepticism, even though there is no persuasive argument from externalism to the falsity of skepticism.
18. An externalist might try to reject this suggestion, and instead continue to maintain that knowledge of content is non-inferential and non-empirical, and that it allows for knowledge of the world without being dependent on knowledge of the world. But, to reiterate and emphasize an earlier point, this does not result in an interesting anti-skeptical argument: Either the externalist premise is left unfounded, in which case the argument can be rejected, or it is supported, in which case the argument begs the question against the skeptic, and can be rejected; in neither case is there a reason to accept the argument.
19. This raises a further problem for Putnamian anti-skeptics: If externalism is indeed inconsistent with privileged knowledge of content, then there is a problem with S' and any other Putnamian anti-skeptical argument of similar form: Its premises would be inconsistent. If externalism is true, we do not have privileged knowledge of content, and if we have privileged knowledge of content, externalism is not true. Given this, the two sorts of premise cannot conspire in an interesting argument against skepticism. Putnamian anti-skeptical arguments, then, would inevitably be unsound.

Notice that this point does not claim that Putnamian anti-skeptical arguments are refuted by first showing that externalism is inconsistent with *a priori* self-knowledge. Rather, it is first shown that externalism is inconsistent with skepticism; the inconsistency of externalism and self-knowledge is then offered as an explanation of the inconsistency of externalism and skepticism. The direction of argument is exactly the reverse of that addressed by Warfield (forthcoming), and is therefore untouched by Warfield's remarks.
20. Certain of Burge's points can be seen to argue against the *reductio* on the grounds that our knowledge of the contents of our thoughts is not of the sort that could conspire with externalism to refute skepticism in the manner of argument S' above. Though we can know the contents of our thoughts, we do not know them to the extent that we can rule out that we have the concept of water as opposed to the concept of twater, or the concept of a brain as opposed to the concept of electronic configuration that is perceptually indistinguishable from a brain. Therefore, I cannot, in premise (A), appeal to my knowledge that I am thinking that I am not a brain in a vat. Burge does not take himself here to be denying what is patently true

about self-knowledge. He does, of course, believe that we have knowledge of content, it is just that this knowledge is not as strong as (A) requires.

I take it, however, that our discussion of the standard strategy in the previous chapter refutes this view of self-knowledge. Thus, I do not think that Burge can develop any resistance to the *reductio* along these lines.

CHAPTER 4

1. To maintain contact with the relevant literature, I will in this part of the book use the terms "anti-individualism" and "individualism" rather than "externalism" and "internalism".
2. I am following tradition in attributing the theory to Marr despite the substantial contributions of others.
3. Segal (1989, p. 199) notices this error in Burge's analysis, and supports it by careful exegesis of Marr's work, but he does not develop it the way I propose to do here. Davies (1991, p. 465) picks up on this point as well, and Egan (1992) also shows sensitivity to it, though in my view she extends the point too far in arguing that *all* talk of content in Marr's theory is heuristic, such that there is no commitment in the theory even to the existence of content (see section 3.3 below and Butler, 1996b, for criticisms of Egan's view).
4. Davies sees Segal's liberal content assignments as part of what he calls a "revisionary individualism", since the contents assigned will not agree with "workaday" content assignments, which utilize descriptions familiar from the language of folk psychology. The "conservative individualist" accepts the workaday content assignments, but maintains that contents still supervene on states internal to the individual.
5. Egan (1991, 1992) has contested this latter claim; I have criticized Egan's position in section 3.3 below and in Butler (1996b).
6. Wilson (1994) has argued that individuation by computational properties is not necessarily individualistic. I argue against Wilson's view in chapter 6.
7. The T-I test is usually taken as a test for individualism insofar as it is a test to determine if visual states, specifically the contents of visual states, supervene locally on states internal to the individual. For our purposes, a psychological state supervenes on an individual's internal states if and only if an individual's mental kinds could not be different unless the individual's internal constitution were different. As Davies (1991, p. 463) points out, however, individualism is not equivalent to local supervenience; in a world where it is metaphysically impossible for an individual's environment to be different unless her internal states were different, "...sameness of internal constitution would suffice for sameness of mental kind, even if the mental kind depended for its individuation upon the external environment." So, merely establishing local supervenience does not by itself establish that individualism is true. It would also have to be shown that there is no metaphysical impossibility in varying the environment without varying an individual's internal states. It is generally taken for granted, however, that there is no such impossibility, so a demonstration of local supervenience is usually counted as good evidence in favor of individualism.
8. Many commentators on Marr's theory, and indeed throughout the literature on the philosophy of cognitive science, perniciously allude to information that must be "represented" by the visual system or some other cognitive system. This confuses a simple distinction between the information (or informational content), which is of or about, e.g., states of affairs in the world, and the states of, say, the visual system (or, possibly, the visual system plus its history of environmental relations), which *encode* that informational content, and thus *represent*, not the information, but the states of affairs in the world. It will help matters considerably to remain clear on this point.
9. Burge (1986a, p. 25) explains his interest in Marr's theory, in part, by appealing to the claim that "the theory provides an example of mentalistic theory with solid

achievements to its credit." Shapiro (1993, p. 504) also appeals to "the amazing success of [Marr's] project" to justify taking seriously the theory's implications.

10. As I mentioned in fn. 2, Egan (1991, 1992) argues that Marr's theory concerns only the computational mechanisms of the visual system; all talk of content is heuristic and belongs to explanatory models of the theory, which help to make the theory intelligible, but not to the theory itself. Shapiro is sharply critical of this claim. While I agree that Egan's claim is incorrect, she is closer to being right than Shapiro appreciates; Marr's discussions of content are used for identification rather than individuation, just not in the way Egan imagines (see sections 3.4 and 5.2 below).

11. I will take it for granted that Egan is right that the computational properties of mechanisms and states in the visual system are determined by factors internal to the system. Some (e.g., Wilson 1994b) have questioned this supposition, in my view without merit (see my discussion of Wilson's views in chapter 6).

12. Though these remarks have been critical of Egan's view of Marr's theory, there is at least one respect in which she appreciates an aspect of Marr's theory that others have overlooked. This is discussed in section 5.2 below.

13. Though Shapiro (1993, p. 489) claims not to be arguing in his paper for the claim that Marr's theory is non-individualistic, he does say explicitly (fn. 4) that he will be discussing parts of Marr's theory that "entail the theory's non-individualism". Moreover, he exploits the claim that the content assignments are non-individualistic in arguing that non-individualistic theories can pass the T-I test, and hence that the T-I test is not diagnostic of individualism (see below). Therefore, there is no distortion in attributing to Shapiro the conviction that Marr's theory is non-individualistic, and that it is non-individualistic for the reasons I have cited in the text.

14. It must be emphasized that the mere fact that the states have representational contents that are about external states of affairs does not make the states with those contents non-individualistic. The position I will be defending is that the states refer to, or are about, external features, but that the states have those contents in virtue of special properties internal to the states themselves. Thus, the contents are at once broad (in the sense that they are about external features) and individualistic (in that they have their contents in virtue of factors internal to the states themselves).

15. Some theories do tell us how to individuate contents. Stampe's (1977) causal theory and Dretske's (1981) information-theoretic theory (and the 1986 version), for example, appeal to relational properties to individuate contents. But there is no reason to suppose that Marr's theory harbors a commitment to such individuative apparatus.

16. Segal (e.g., 1989, p. 192) seems to drop this epistemological reading of Marr when he claims that, according to Marr's theory, the visual system must itself make assumptions about the distal environment in inferring the 3-D model representation from its impoverished gray level arrays. But, Egan (1992, p. 449) notices correctly that the assumptions are external to the system, and simply made by the theorist in determining what contents to assign at the various stages of processing. The visual system itself makes no assumptions, but merely "...operates in such a way that if the assumptions are true it will succeed in recovering information about the environment from information in the input." Given that Segal is not always careful with his own distinction, it is perhaps too harsh on Shapiro to register surprise that he does not exploit the distinction properly.

17 In fairness to Shapiro, he does acknowledge that Marr has no interest in philosophical attempts to *naturalize* content. Nevertheless, he does seem to think Marr has an interest in committing himself to a philosophical view about how contents are individuated into kinds. It is this latter supposition that I think Shapiro is mistaken about.

18. So, Shapiro (1993, p. 510, fn. 51) overstates his complaint that Egan is guilty of "a deep misunderstanding of Marr's theory" (Shapiro has, in conversation, conceded as much). Though it is true that Marr does not eschew content from his

theory, as Egan suggests, neither does he embrace a metaphysical theory of content (e.g., of a causal or information-theoretic variety).

19. Shapiro also criticizes Segal's claim that there is nothing in Marr's theory that would motivate assigning different contents in a thought experiment involving twins. In the section 6, we will visit this aspect of Shapiro's critique of Segal more closely.

20. For example, consider one passage where Burge talks of individuating visual representations:

> "...their specific intentional content *depends* partly on their being the normal products of the specific objective entities that give rise to them. That is why we *individuate* intentional visual representations in terms of the objective entities that they normally apply to..." (Burge, 1986a, pp. 40-1, emphasis added).

In this passage Burge claims that contents "depend" on their normal distal causes, and that this provides a justification for the claim that we "individuate" contents in terms of their normal distal causes. I think it is clear from the context surrounding this passage, and from the passage itself, that Burge's use of the term "depends" serves to indicate that content types are metaphysically fixed partly by their normal distal causes. But if this is so, and it provides us with, as he puts it, a reason for "why we *individuate* intentional visual representations in terms of the objective entities that they normally apply to", then it cannot be that this use of the term "individuate" has anything to do with metaphysically fixing the content kind. For if it did, then Burge would be offering as a reason for individuating contents a certain way the very fact that we individuate them in that way (as, for example, when Jesse Jackson was reported to have said, "All of us cannot be famous, because all of us cannot be well known." (Quoted by Elizabeth Drew in "A Political Journal", which appeared *The New Yorker*, March 12, 1984, p. 140.) Clearly the one claim simply restates the other, and is therefore not a reason for it. It would be gratuitously uncharitable to suppose that this is all that Burge is doing in the quoted passage. Therefore, either the claim that contents depend on distal causes is not a metaphysical claim, or the claim that contents are individuated by us in terms of distal causes is not a metaphysical claim. I think it is plausible to assume that it is the latter claim that is not metaphysical, and should instead be interpreted to be the claim that we *identify* contents in terms of distal causes. Indeed, even the way Burge puts the point, so that *we* do the individuating, suggests that it is not intended to be a metaphysical claim; individuation, properly and metaphysically understood, concerns how the visual system divides into kinds, not how we might divide it into kinds.

21. This should not be construed as an attempt to defend Shapiro's criticisms of Segal; I think, as a matter of fact, that even once the metaphysical interpretation is exported several deep difficulties remain with Shapiro's criticisms of Segal. First, Shapiro (1993, p. 506) attributes to Segal the view that content assignments are derived from discriminative behavior. As a matter of fact, that is only one constraint on content assignments, according to Segal. Second, Shapiro (1993, p. 508) claims to show that discriminative behavior is *irrelevant* to constraints on content assignments. However, he argues (1993, p. 506) only that it is *insufficient* to constrain content assignments fully, and then slides without acknowledgment to the stronger irrelevancy claim. Third, insofar as the appeal to the computational theory is a sort of top-down constraint on content assignments, it is no particular advance upon Segal's own view, despite Shapiro's (1993, p. 509) advertisements to the contrary.

22. Segal might wish here to appeal to the modular character of vision, and claim that constraints from outside a module are irrelevant (see, e.g., Segal's, 1991, response to Davies). I do not have a clear refutation of such a move. Nor, however, is there a clear reason to endorse it. Moreover, the general problem of matching visual contents and attitude contents still remains (see below).

23. I will note, however, that Burge misunderstands in the third premise what is required for a representation to be empirically informative. He claims (1986a, p. 40) that

> "...some visual representations that represent objective entities as such must have the representational characteristics that they have [i.e., their content] partly *because* instances regularly enter into certain causal relations with those objective entities."

As a matter of fact, it is sufficient that representational states have the representational content (type) that they do owing to factors internal to the individual (and hence individualistically). We need only assume that evolution has selected a system that reliably deploys representations with content type X when confronted with Xs; in this case, Xs are the normal distal causes of the *tokens* of that representational type, even though that representational type is determined to be the type that it is by internal features, independently of the normal distal cause of tokens of that type. Burge assumes, without argument, that it is the normal distal causes, the Xs, that bestow the type individuation on the tokens they cause. I claim that the reverse is just as plausible; that the Xs, the normal distal causes of those tokens, are in fact that normal distal causes because the tokens caused already are type-individuated as representations of Xs, and it is adaptively advantageous to deploy representation of Xs when confronted with Xs.

Burge says, immediately following this discussion, perhaps in support of his claim, that "this is the core of truth in the slogan, sometimes misapplied ... that mistakes presuppose a background of veridicality". I disagree. The core truth in the slogan is that it makes no sense to talk of a mistake except as a deviation from veridicality. If there were no such thing as veridicality, there could be no such thing as a mistake. It is not, as Burge suggests, that our making mistakes *shows* that we are right most of the time. It may be true that we are right most of the time, but this need not be the case in order for there to be mistakes, nor in order for us to make sense of the fact that we make mistakes.

24. Segal is right, however, to criticize Davies and Egan for supposing that liberal contents must be highly disjunctive (see Segal, 1991, p. 490; this is his second VIP). Liberal contents are only disjunctive relative to restrictive individuation of the kinds in the extension of the content. Segal sometimes calls his content assignments "narrow", insofar as they supervene on the internal features of the subject. They are *not* narrow, however, in the sense of not determining an extension. It is clear that Segal takes the liberal contents to determine an extension, just one that is more inclusive (but no more disjunctive) than that determined by the restrictive contents.

25. Interestingly, Davies (1991, p. 467) quotes the same passage from Burge as evidence in the case against the original defense of conservative individualism. However, he leaves off the italicized phrase, the very phrase that contains the point on which I am focusing. In part for this reason, he is not well-positioned to anticipate the possibility of a counterfactual case where Twin P is well-adapted to an environment that is relevantly different from P's environment.

26. Notice that this is not the charge that Shapiro (1993, p. 494) levies against Burge. Shapiro assumes that Burge is taking Marr's theory to have direct individuative implications, and that Burge begs the question against individualism by assuming that those implications are individualistic. I am suggesting that Burge is merely taking Marr's theory to *assign* contents, but that in assuming that the theory would be *correct* in assigning different contents to the twins, he must presuppose that individualism is false (because if it is true, then Marr's content assignments in the counterfactual case would be false).

CHAPTER 5

1. The list includes Burge (1986b, 1989a), Egan (1991), Wilson (1992, 1993), and McClamrock (1995), among others.

2. Fodor (1991) contains developments of this basic argument not included in his (1987, ch. 2). These will be discussed briefly in section 4.
3. Burge raises but does not pursue difficulties with the first three premises: (1) and (2) ignore Davidson's claims concerning the individuation of psychological states by appeal to their causal relations. Also, the relevant causal relations need not be effects; causes are causally relevant as well. (3) fails to recognize that an anti-individualist might individuate *brain* states by appeal to environmental relations as well. In spite of these concerns, Burge is prepared to concede (1)-(3).
4. Pete Mandik has suggested the following way to view thought experiments, which I like very much. Thought experiments are essentially equivalent to computer simulations. Such simulations take the place of real experiments by taking us from a formal representation of certain initial conditions of a system to a formal representation of later conditions of the system according to constraints taken from empirical laws and antecedent metaphysical commitments. Thought experiments are simply "neck-top" computer simulations.
5. There are actually at least two *prima facie* problems in understanding the causal relevance of intentional properties. The first is now being presented in the text, and the second involves the anti-individualists' claim that intentional content is relationally individuated. The second will be the subject of the remainder of the chapter. I argue that it is incompatible with the causal relevance of the intentional that the intentional be relational. Thus, our sought-after motivation for a plausible individualism.
6. See, e.g., Davidson (1980), Lewis (1966), and Kim (1993). For refinements on this closure principle, see Yablo (1992) and Robb (1997). I will not pursue these and other subtleties since my aim is to sketch only the broad contours of a solution to the puzzle, and then illustrate how anti-individualism is inconsistent with any solution that conforms to these broad contours.
7. Various distinctions can be drawn between varieties of supervenience (see, e.g., Kim, 1993). One must distinguish global from local supervenience (i.e., whether the relation holds between the total set of physical and mental properties, or just between distinct (or distinct subsets of) physical and mental properties), and strong from weak supervenience (i.e., whether the dependency is characterized by appeal to all logically possible worlds or just those worlds that are nomologically possible). The proposed solution I will be discussing is concerned with local supervenience, and is agnostic between the stronger and weaker forms.
8. Heil (1992, ch. 4) presents a refinement of the supervenience solution that is similar to Robb's. I will discuss Heil's treatment at some length in section 2.4 of this chapter).
9. Robb (1997, pp. 188-190) illustrates how the trope solution entails supervenience of the sort discussed above. Thus, it is not so much that supervenience is irrelevant to an account of mental causation as it is that supervenience does not provide the whole solution.
10. Robb (1997, pp. 190-194) entertains several sorts of objection to the trope solution to the problem of mental causation, one of which is skepticism about tropes in the first place. Thus, one might claim that the mystery of mental causation has been replaced by the mystery of tropes. But, as Robb points out, tropes have been posited for reasons entirely independent of the mental causation issue, and may even derive some abductive support from the fact (if it is one) that they provide for the best (or only?) solution to the mental causation puzzle.
11. This is not to say that content would have to fail to make reference or be truth-conditional; it is just to say that the content would have whatever intentional properties (tropes) it has solely in virtue of the physical tropes with which it is identical.
12. Thanks to Dave Robb for useful electronic discussion on the implications for anti-individualism of a trope construal of mental causation.
13. Wilson claims (in correspondence) to have no interest in defending Burge's notion of wide causal powers, but he does (e.g., 1992, p. 120) draw a distinction between intrinsic or restricted (non-relational) causal powers and extrinsic or

extended (relational) causal powers that might be pressed into service on behalf of the anti-individualist. It is this sort of notion that I will be challenging in this chapter.

14. I note in passing that there are reasons to think that the multiple realizability of psychological states by neural kinds is often overstated; see Butler (1994a) for arguments to this effect. Also, there are reasons for thinking that neural states are not context-sensitive from the perspective of psychology; see Butler (1995a) for further support and elaboration of this point.

15. Evolutionary biology might seem to be an exception insofar as species are type-individuated in terms of the history of a given organism. However, the remarks beginning in the next paragraph, as well as those in the next section, tell against such an appearance.

16. So far as I can tell, Salmon was not himself concerned with the issue of different levels of causal organization. Indeed, his talk of "*the* causal structure of the world" (1977, p. 162; 1984, p. 19) suggests that he took there to be a single level of causal organization, a "causal nexus", into which all events must fit.

17. My suspicion is that Fodor's infamous H- and T-particle argument was designed to illustrate how certain kinds of relation can be screened off. Burge is right that there are crucial disanalogies between H-particles and wide content. But Fodor is surely right in thinking (if he was thinking this) that the relational components in both can be screened off from the effect. Moreover, as we will discuss in a later section, Burge is *not* right when he responds to the H- and T-particle argument by suggesting that psychology is free from the individuative constraints of physics, and may postulate causal powers *independently* of the postulation of causal powers in physics.

18. As we will see below, this point is equally effective against Heil's (1992, ch. 4) discussion of causal powers, as well as that of Macdonald (1989).

19. As I mentioned above, I dispute this claim below.

20. There is a further disanalogy between psychology and the special sciences Burge and the others cite. Take Burge's example of the heart and its homologue that pumps waste. The example is not particularly similar to the Twin cases in psychology in that the psychological cases involve subtle differences in environmental features (i.e., aluminum vs. twalum) that have no causal bearing on the individuals, whereas in the case of the homologous heart, the environmental variations (i.e., blood vs. waste) are bold and dramatic. If instead of waste we consider an homologous organ that pumps a microstructurally distinct sort of blood-like substance (that is largely indistinguishable from real blood), it is not so clear that the organ would be individuated relevantly differently by biology; it may, in other words, just be a heart that pumps a different sort of blood. To the extent that hearts are individuated relationally, then, they are relational in rather unlike the way that Burge takes psychological states (or mechanisms) to be relational. Perhaps psychological mechanisms could be individuated relationally, just insofar as they must operate on representations and not, say, ion clusters; a brain would not be a mind if it took ion clusters as inputs and outputs. But Burge surely wants a stronger claim.

21. Saidel might object to this characterization on the grounds that anti-individualists are not *compelled* to accept wide causal powers; rather, such a move is *open* to them. This objection to my characterization, however, would be disingenuine since Saidel himself is urging anti-individualists to make this move to avoid the tension exposed by Owens. Thus, anti-individualists are not compelled to accept wide causal powers only to the extent that they are not compelled to avoid Owens' tension.

22. We will look at Fodor's suggestion more carefully in section 4 of this chapter.

23. I leave aside the fact that the Twin Earth thought experiments *themselves* argue for nothing; it is only the particular way in which they are interpreted and exploited that argues for wide content. One can therefore entertain the Twin Earth thought experiments, and interpret them in ways that do not militate in favor of wide content.

24. It is, of course, obvious that variations in internal states results in variations in effects, and hence variations in causal powers, even where current (and historical) contexts are held fixed. But this case is irrelevant to the supervenience of causal powers on internal states, which involves only the question of variations in effects *without* variations in internal states.
25. It is important to note, however, that this in no way entails an interest-relativity of causal powers. Interests determine which causal powers we attend to; they do not determine *anything* about causal powers themselves. Wilson comes uncomfortably close to making this type of error in claiming that "the 'pragmatics' of explanation has implications for our metaphysics" (Wilson, 1993, p. 70). Objective probability theory may have implications for our metaphysics through such notions as screening off; there is no warrant for such claims on behalf of pragmatics.
26. On Burge's behalf, this point helps us to see why he is *right* to reject an argument of Fodor's to the conclusion that psychology must taxonomize states in the same way as neurophysiology. Different sciences partition the space of causal powers differently, and can therefore attend to different causal powers of the same token. There is no anomaly in supposing that psychology attends to different causal powers of a token brain state than neuroscience. Of course, neither of these sciences will be free to consider causal powers that are independent of the causal powers recognized by physics (see next paragraph). Thus, even granting the frailty of this particular Fodorian argument, there is no reason to suppose that psychology will taxonomize such states in the way Burge would like.
27. This, then, is why Burge is wrong to think that the causal powers of psychology may be independent of the causal powers of physics. The appearance of higher levels of causal organization, we have seen, may introduce causal powers not present at the physical level, but neither are these causal powers independent of the causal powers of physics. In addition, we have seen that higher causal levels do not offer the possibility of introducing more context-dependence of causal powers, so there is no reason why an example of screening off from physics (e.g., the H- and T-particle story) should not be relevantly analogous to an example of screening off in psychology (e.g., wide content).
28. The control for contexts, in this case, is not possible in actual experimentation. That is why we need philosophical thought experiments to do metaphysics; thought experiments help us to see what would happen in experiments that in actual fact may be impossible. This is also why science is not metaphysics, and why appeals to scientific practice in the service of metaphysical claims, such as anti-individualism, must be interpreted carefully, not superficially.
29. Notice, again, that Twin S's mental state can retain its history of environmental relations, and hence be the same anti-individualistically individuated mental state in the Earthly context. Thus, it is false to suppose, as Saidel (1994) does, that a test like CT assumes an individualistic individuation of mental states.
30. Because Adams frequently collaborates in print with his colleagues, and because there are so many colleagues with whom he collaborates, I will refer to them collectively by the easily remembered short-hand 'the Adams Family'.
31. It would be another thing to argue that distinguishing behaviors in this way is relevant to the interests of psychology. The Adams Family argues (1991, p. 46) that psychology is as interested in explaining why people intentionally drink water rather than twater for just the same reasons that it is interested in explaining why people intentionally drink liquor rather than tea. But this is clearly a bad argument. People can distinguish between liquor and tea; that is why psychology is interested in explaining why people intentionally do one rather than the other. But people cannot, by hypothesis, distinguish between water and twater. For that reason, any explanation of why one drinks water would suffice to explain why one drinks twater. The same is clearly not true in the case of liquor and tea.
32. I should note that Dretske is aware of this problem, and attempts to face up to it. Interested readers should consult Dretske (1988) for discussion.

33. Notice, however, that the failure of local supervenience does make such irrationality possible in a way that is implausible. Imagine a slow switching case of the sort discussed in Burge (1988b), where one is transported to Twin Earth surreptitiously. Suppose one was reasoning in a very protracted way about water. After many years on Twin Earth, one would no doubt be associating one's term "water" with twater; so, by anti-individualist lights, one's concept of water would have mutated into a concept of twater. But this would not affect the inferences one would draw in one's protracted reasoning. An inference that began on Earth but was completed many years later on Twin Earth would involve an equivocation on formulae involving the term "water"; early in the reasoning it would mean water, later it would mean twater. That such equivocation is even possible strikes me as quite counterintuitive. And given the prevalence of slow switching cases, as I argued in ch. 2, this counterintuitive irrationality would be quite prevalent if anti-individualism were true. This is not a knock-down argumentagainst anti-individualism, of course, but it can plausibly be taken as a strike against it, one that would motivate the search for a plausible individualistic account of mental states.
34. Burge, of course, rejects metaphysical theorizing in favor of a more transcendental view of causal powers. He suggests that we take the "understanding" of causality yielded by explanations in the special sciences, specifically psychology, as the primary data for identifying causal powers (see, e.g., Burge, 1993, p. 103). Though Burge is offering this in the context of questions about epiphenomenalism and the possibility of mental causation, it is clear that such sentiments lie behind his willingness to let psychology and other special sciences fix their space of causal powers independently of physics (see, e.g., Burge, 1989a, p. 318). There are at least two problems with this suggestion, however: It is not clear that we *have* an antecedent understanding of mental causation on which to rest our view of causal powers, and even if we did, there is every reason to think that it should be overridden by the arguments against relational causal powers advanced in this chapter.

CHAPTER 6

1. Fodor and Pylyshyn (1988, p. 9) appear to make such a claim when they identify representation as the ingredient necessary for a system to exhibit a cognitive level of organization.
2. I am not here begging the question by assuming that the relevant features of the system in virtue of which it is representational are internal features. It could be that the relevant features are relational. In fact, Dretske (1988, ch. 3) distinguishes between representational systems that derive their intentionality from use by others, and those that represent intrinsically; he maintains, however, that the relevant intrinsic properties are historical and relational in nature. I take the arguments of this book to argue indirectly against Dretske's information-theoretic psychosemantics. But I am in agreement with him that there is a distinction to be drawn between intrinsic and derived intentionality.
3. The claim that psychological states such as beliefs and desires explain only intentions should not be confused with Hornsby's (1980) claim that behavior, or action, just is the intention or trying. I am content to accept the ordinary, perhaps non-autonomous, descriptions of behaviors as public movements. I maintain only that various contextual assumptions must be made in order to explain behaviors under these descriptions; these contexts, however, are non-psychological, and there should be no temptation to claim that everything cited in a behavioral explanation is psychological. Psychological states can explain, on their own, only further psychological states.
4. Even if the anti-individualist were to counter this move, and argue that psychological explanations are explanations of behavior (perhaps under appropriate interpretation), rather than psychological phenomena such as intentions, it is not likely that they can charge that the individualist is left with

nothing to explain. It is possible, in other words, to undercut the plausibility of the claim that autonomous descriptions of behavior are difficult to come by. Consider a suggestion from Walker (1990). Walker is reacting to an argument of Owens (1987) that makes, in the context of a larger discussion, a point very similar to the one expressed above. As Walker in effect notes, autonomous behavioral descriptions are hard to come by only so long as we look at descriptions couched in context-involving language. Walker's suggestion is to start with behavioral descriptions of this sort, and peel away their context-infested, non-autonomous aspects:

> "First, we individuate behavior the way common sense does, in terms of folk intentional action descriptions. ... We can then group all actual and possible instances of physical behavior into sets according to these intentional descriptions. So, any common-sense intentional description of behavior D has as its extension the class of actual and possible actions, viz., those actions that comprise all the actual and possible referents of D.
>
> Next we need to include *all* those instances of actual and possible behavior functionally identical to the members of D's extension, viz., the relevant instances of Doppelganger behavior. ... Call this new set of corresponding Doppelganger-behavior Doppel(D). ...Doppel(D) is the set of all actual and possible instances of behavior that are physically identical to some member of D but are performed by a Doppelganger of the agent of action. ...
>
> The extension of D plus the extension of Doppel(D) together now include *all* the instances of actual and possible behavior that are functionally equivalent -- the union contains exactly the behavior [individualistic] psychologists want.... (Walker, 1990, pp. 429-30.)

With Walker's recipe in hand, the individualist need not fret over the many examples of non-autonomous behavioral descriptions that are available in ordinary language. The psychological relevance of any one of them can be given an autonomous description in terms of D and Doppel(D). Thus, even if we were to admit that individualists were in need of autonomous behavioral descriptions, it would simply not be true that they lack a way of describing their version of psychological explananda.

5. It is also worth mentioning two further points: First, this line of argument overestimates the plausibility of multiple realizability in folk psychology and cognitive psychology (see Butler, 1994a, for extended criticisms of arguments for multiple realizability). Second, this line of argument appears to assume a principle according to which sameness of behavior entails sameness of causal explanation. Obviously, such a principle is implausible on its face and should be rejected.

6. Wilson (1994a, p. 57) seems to recognize this point, but there does not appear to be any attempt to respond to it.

7. One might also worry that the restriction to nearby worlds, or relevant alternatives, might just as easily be brought in in responding to Twin Earth thought experiments in application to the meaning of psychological states. Since Twin Earth is not a relevant alternative, there is no reason to consider how the identity of psychological states might change if one were an inhabitant of Twin Earth.

As I indicated in chapter 2, however, I think there are plenty of relevant "twin" situations. I think the same point can apply here; there are plenty of things the so-called face recognizer could respond to that are not faces, so it is not true that the thing is *just* a face recognizer, as Wilson claims.

8. One way it could turn out that the functional individuation of the device would not make any essential appeal to the distal environment is if mental contents turned out to supervene on internal states. In that case, all contentful states, including states of the face recognizer, would be individuated internally. The description of the device a face recognizer, then, would be simply an expository convenience that

had no individuative significance. If the world contained no faces, ever, that would not affect the role the "face" recognizer plays in one's cognitive economy.
9. There is the further question about determining the boundries of the system relative to which certain aspects are considered external while others are internal. This will be the focus of concern in section 4 below.
10. There are also serious questions to be raised about the viability of the so-called "evolutionary psychology" of Cosmides and Tooby to which Wilson (1994a, p. 70-71) appeals (see, e.g., Cosmides and Tooby, 1987). Cosmides and Tooby's work aims to establish the psychological relevance of evolutionary considerations by defending a 'social exchange theory of reasoning'. The social exchange theory claims that humans sometimes reason by means of content-dependent, domain-specific algorithms that are the inevitable consequence of our selectional history. Davies, Fetzer, and Foster (1994), however, have argued, quite convincingly, that evidence in support of social exchange theory is entirely absent.
11. An anonymous reviewer has suggested that other work in developmental psychology might be brought in to save Wilson's point. In particular, Karmiloff-Smith's (e.g., 1992) work on child development and Rutkowska's (e.g., 1993) treatment of 'the computational infant' have been taken by some (e.g., Losonsky, 1995) to provide key evidence in favor of anti-individualism. The nature of the anti-individualism, however, appears to be much more akin to systemic anti-individualism. For that reason, I will discuss this work in section 4 of this chapter, which is devoted to the issue of systemic anti-individualism.
12. In fairness to Dretske, it may be that he is assuming a process account of behavior, even though he does not say so explicitly. In that case, the explanation of the process may be anti-individualistic. Alternatively, he may be taking his appeal to explanatory practice to motivate, or perhaps even justify, a process view of behavior. In either case, however, the resulting anti-individualistic explanation would appear to be a structuring cause (how the triggering cause of certain movements came to cause those movements), rather than a triggering cause. Since our concern has been with triggering causes (and since Dretske is himself claiming to be concerned with triggering causes), neither of these alternatives will help in the present context.
13. See, e.g., Toumela (1989), Haugeland (1993), Port and van Gelder (1995), Clark and Chalmers (in preparation), and Wilson (1994b) who makes this point in the context of computational accounts of cognition.
14. A philosophically oriented discussion of such feedback systems, or servo systems, can be found in Butler (1992). A psychologically oriented discussion can be found in Gallistel (1980, pp. 275ff). An engineering-oriented discussion can be found in Merton (1973). See also the discussion in the following section.
15. The appeal to Reichenbach's common cause principle is perhaps better seen as arguing that, given certain external contextual assumptions (i.e., that there is a tree), the object of the perceptual representations is the tree (a distal object) rather than the various retinal images (proximal events). This, in fact, is the suggestion of Forster (in preparation).
16. An anonymous reviewer has indeed made this objection.
17. It is quite natural to complain that Brooks' subsumption architecture is too simple to deal with the complexities of human existence, and so is an inappropriate model of human cognition (see, e.g., Kirsh 1991). Though I think this is true, I do not begrudge Haugeland's attempt to make his case in increments. It would be bad enough, from my point of view, if Haugeland were right about ants and robots. I am therefore duty-bound to challenge his points even at this level.
18. Servomechanisms are control structures that compare sensory signals to a stored representation (efference copy) and measure the difference between them (there are well-specified dimensions along which differences can be measured). Detected differences result in an error signal being sent to muscle groups whose activity will lessen the amount of difference between the sensory signal and the efference copy. This is the principle means of navigation in power-steering systems

and guided missiles. See (Merton 1973) for a readable introduction to servomechanisms.

19. Massive environmental anomalies, of course, can corrupt the smooth flow of information between the environment and lower-level servomechanisms, and call for reassessments from above. But this is just what we would expect from a mind that is *largely* but not totally insulated from its environment; it presents no reason whatsoever to wash away the distinctions that separate mind, body, and world.

20. It is worth pointing out that Dretske's more recent writings (and conversations) indicate that he no longer expects the multiple pathways solution to carry the burden in explaining why events in the distal environment are the objects of representation rather than more proximal events. Thus, the criticism in the text attaches only to McClammrock's view, which has been advanced some 15 years after Dretske's.

REFERENCES

Adams, F. (1991) "Causal Contents", in McLaughlin (1991), pp. 131-156.

Adams, F. (1993) "Fodor's Modal Argument", *Philosophical Psychology*, 6, pp. 41-56.

Adams, F. and Fuller, G. (1992) "Names, Contents, and Causes", *Mind and Language*, 7, pp. 205-221.

Bach, K. (1987) *Thought and Reference*. Oxford: Clarendon Press.

Bach, K. (1988) "Burge's New Thought Experiment: Back to the Drawing Room", *Journal of Philosophy*, Vol. LXXXV, pp. 88-97.

Ballard, D. (1991) "Animate Vision", *Artificial Intelligence*, 48, pp. 57-86.

Bernecker, S. (1996) "Externalism and the Attitudinal Component of Self-Knowledge", *Nous*, 30, pp. 262-275.

Bilgrami, A. (1992) "Can Externalism Be Reconciled with Self-Knowledge?", *Philosophical Topics*, 20, pp. 233-267.

Bilgrami, A. (1994) *Belief and Meaning*. Oxford: Basil Blackwell.

Boghossian, P. (1989) "Content and Self-Knowledge", *Philosophical Topics*, 17, pp. 5-26.

Bonjour, L. (1991) "Is Thought a Symbolic Process?", *Synthese*, 89, pp. 331-352.

Brandon, R. (1984) "The Levels of Selection", in R. Brandon and R. Burian (eds.) *Genes, Organisms, Populations: Controversies Over the Units of Selection* (pp. 133-41). Cambridge, MA: MIT Press.

Brooks, R. (1991), "Intelligence Without Representation", *Artificial Intelligence* 47: 139-159.

Brown, J. (1995) "The Incompatibility of Anti-Individualism and Privileged Access", *Analysis*, 56, July.

Brueckner, A. (1986) "Brains in a Vat", *Journal of Philosophy*, 83, pp. 148-167.

Brueckner, A. (1990) "Scepticism about Knowledge of Content," *Mind*, 99, pp. 447-52.

Brueckner, A. (1992a) "If I Am a Brain in a Vat, Then I Am Not a Brain in a Vat", *Mind*, 101, pp. 123-128.

Brueckner, A. (1992b) "What an Anti-Individualist Knows A Priori", *Analysis*, 52, pp. 111-18.

Brueckner, A. (1994) "Knowledge of Content and Knowledge of the World", *The Philosophical Review* 103 (1994): 327-343.

Burge, T. (1979) "Individualism and the Mental", *Midwest Studies in Philosophy*, Vol. IV, pp. 73-121.

Burge, T. (1982) "Other Bodies", in Woodfield, A. (ed.), *Thought and Object.* Oxford: Clarendon Press.

Burge, T. (1986a) "Individualism and Psychology", *The Philosophical Review,* 95, pp. 3-45.

Burge, T. (1986b) "Cartesian Error and the Objectivity of Perception," McDowell, J. and Pettit, P. (eds) *Subject, Thought and Context.* Oxford: Clarendon Press.

Burge, T. (1986c) "Intellectual Norms and Foundations of Mind", *Journal of Philosophy,* Vol. LXXXIII, pp. 697-720.

Burge, T. (1988a) "Authoritative Self-Knowledge and Perceptual Individualism", in R. Grimm and D. Merrill (eds.) *Contents of Thought.* Tucson: Arizona University Press.

Burge, T. (1988b) "Individualism and Self Knowledge", *Journal of Philosophy,* 85, pp. 649-63.

Burge, T. (1989a) "Individuation and Causation in Psychology", *Pacific Philosophical Quarterly,* 70, pp. 303-322.

Burge, T. (1989b) "Wherin is Language Social?", in A. George (Ed.) *Reflections on Chomsky.* Oxford: Basil Blackwell.

Burge, T. (1993) "Mind-Body Causation and Explanatory Practice", in J. Heil and A. Mele (eds.) *Mental Causation.* Oxford: Clarendon Press.

Butler, K. (1992), "The Physiology of Desire", *Journal of Mind and Behavior* 13: 69-88.

Butler, K. (1994a) "Neural Constraints in Cognitive Science", *Minds and Machines,* 3, pp. 129-161.

Butler, K. (1994b) "The Scope of Psychology", *PSA Vol. 1,* pp. 428-437.

Butler, K. (1995a) "Compositionality in Cognitive Models: The Real Issue", *Philosophical Studies,* Vol. 78, pp. 125-151.

Butler, K. (1995b) "Content, Context, and Compositionality", *Mind and Language,* Vol. 10, pp. 3-24.

Butler, K. (1996a) "Content, Causal Powers, and Contexts", *Philosophy of Science,* Vol. 63, pp. 105-114.

Butler, K. (1996b) "Content, Computation, and Individualism in Vision Theory", *Analysis,* Vol. 56.3, pp. 146-54.

Butler, K. (1996c) "Individualism and Marr's Computational Theory of Vision", *Mind and Language,* Vol. 11, pp. 313-337.

Butler, K. (forthcoming-a) "Externalism, Internalism, and Knowledge of Content",*Philosophy and Phenomenological Research.*

Butler, K. (forthcoming-b) "Externalism and Skepticism", *Dialogue: The Canadian Philosophical Review.*

Butler, K. (forthcoming-c) "Content, Computation, and Individuation", *Synthese.*

Butler, K. (forthcoming-d) "Problems for Externalism and *A Priori* Arguments of Skepticism", *Dialectica.*

Butler, K. and Shogenji, T. (in preparation) "Compositionality in Thought and Language".

Campbell, K. (1990) *Abstract Particulars.* Oxford: Basil Blackwell.

Carlson, N. (1986) *The Physiology of Behavior.* Boston: Allyn and Bacon, Inc.

Casati, R. and Dokic, J. (1991) "Brains in a Vat, Language, and Metalanguage", *Analysis,* 51, pp. 91-93.

Collier, J. (1990) "Could I Conceive Being a Brain in a Vat?", *Australasian Journal of Philosophy,* 68, pp. 413-19.

Chalmers, D. (1996) *The Conscious Mind.* Oxford: Oxford University Press.

Chalmers, D. (forthcoming) "The Components of Content".

Churchland, P.M. (1981), "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy* 78, 67-90.

Clark, A. (1996) *Being There.* Cambridge, MA: MIT Press.

Clark, A. and Chalmers, D. (in preparation) "The Extended Mind", MS.

Cosmides, L. and Tooby, J. (1987), "From Evolution to Behavior: Evolutionary Psychology as the Missing Link", in J. Dupré (ed.), *The Latest on the Best.* Cambridge, MA: MIT Press, pp. 277-306.

Crane, T. (1991) "All the Difference in the World", *The Philosophical Quarterly,* Vol. 41, pp. 1-25.

Cummins, R. (1975), "Functional Analysis", *Journal of Philosophy* 72, pp. 741-760.

Cummins, R. (1989) *Meaning and Mental Representation.* Cambridge, MA: MIT Press.

Darley, J., Glucksberg, S., and Kinchla, R. (1991) *Psychology.* Englewood Cliffs, NJ: Prentice Hall.

David, M. (1991) "Neither Mentioning 'Brains in a Vat' nor Mentioning Brains in a Vat Will Prove that We are not Brains in a Vat", *Philosophy and Phenomenological Research,* 51, pp. 891-896.

Davidson, D. (1980) *Essays on Actions and Events.* Oxford: Clarendon Press.

Davidson, D. (1987) "Knowing One's Own Mind", *Proceedings of the American Philosophical Association,* 60, pp. 441-458.

Davies, D. (1995) "Putnam's Brain-Teaser", *Canadian Journal of Philosophy,* 25, pp. 203-228.

Davies, M. (1991) "Individualism and Perceptual Content", *Mind,* Vol. C. pp. 463-484.

Davies, M. (1994) "Externalism, Architecturalism, and Epistemic Warrant", invited symposium at the Eastern Division Meeting of the APA, Dec, 1994.

Davies, P., Fetzer, J., and Foster, T. (1994), "Logical Reasoning and Domain Specificity", *Biology and Philosophy,* 9, 000-000.

Dell'Utri, M. (1990) "Choosing Conceptions of Realism: The Case of Brains in a Vat", *Mind*, 99, pp. 79-90.
Dennett, D. (1978) "Where Am I?", in *Brainstorms.* Cambridge, MA: Bradford Books.
Dennett, D. (1987) *The Intentional Stance.* Cambridge, MA: MIT Press.
Donnellan, K. (1966) "Reference and Definite Descriptions", *Journal of Philosophy.*
Dretske, F. (1981) *Knowledge and the Flow of Information.* Cambridge, MA: MIT Press.
Dretske, F. (1986) "Misrepresentation", in Bogdan, R. (ed) *Belief.* Oxford: Oxford University Press.
Dretske, F. (1988), *Explaining Behavior.* Cambridge, MA: MIT Press.
Dretske, F. (1992) "What Is Not Wrong With Folk Psychology", *Metaphilosophy*, 23, 1-13.
Dretske, F. (1993) "The Nature of Thought", *Philosophical Studies.*
Dretske, F. (1994) "Introspection", Proceedings of the Aristotelean Society.
Dretske, F. (1995) *Naturalizing the Mind.* Cambridge, MA: MIT Press.
Dreyfus, H. (1972, 1979, 1992), *What Computers Can't Do.* New York, NY: Harper and Row (first two editions only). *What Computers Still Can't Do.* Cambridge, MA: MIT Press (third edition).
Egan, F. (1991) "Must Psychology Be Individualistic?", *The Philosophical Review*, 100, pp. 179-203.
Egan, F. (1992) "Individualism, Computation, and Perceptual Content", *Mind,* Vol. 101, pp. 443-459.
Egan, F. (1994) "Individualism and Computation in Vision Theory", *Analysis*, 51.
Egan, F. (1995) "Content and Computation", *The Philosophical Review*, 104, 181-203.
Eells, E. (1982) *Rational Decision and Causality.* Cambridge: Cambridge University Press.
Evans, G. (1982) *The Varieties of Reference.* Oxford: Oxford University Press.
Falvey, K. and Owens, J. (1994) "Externalism, Self-Knowledge, and Skepticism", *The Philosophical Review* 103 (1994): 107-137.
Fodor, J. (1975) *The Language of Thought.* New York: Crowell.
Fodor, J. (1980), "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology", *The Behavioral and Brain Sciences* 3, 63-73.
Fodor, J. (1983) *The Modularity of Mind.* Cambridge, MA: MIT Press.
Fodor, J. (1987), *Psychosemantics.* Cambridge, MA: MIT Press.
Fodor, J. (1991), "A Modal Argument For Narrow Content", *Journal of Philosophy,* 88, pp. 5-26.

Fodor, J. (1993) "Concepts: A Potboiler", *Cognition*, X, pp. yyy-yyy.
Fodor, J. (1994) *The Elm and The Expert.* Cambridge, MA: MIT Press.
Fodor, J. and Pylyshyn, Z. (1988) "Connectionism and Cognitive Architecture", *Cognition*, 28, pp. 3-71.
Forbes, G. (1995) "Realism and Skepticism: Brains in a Vat Revisited", *Journal of Philosophy*, 92, pp. 205-22.
Forster, M (in preparation) "In Defense of Causal Theories of Meaning", ms, University of Wisconsin-Madison.
Gallistel, C. (1980) *The Organization of Behavior.* Hillsdale, NJ: Erlbaum.
Gallois, A. (1992) "Putnam, Brains in a Vat, and Arguments for Skepticism", *Mind*, 101, pp. 273-86.
Ghez, C. (1985), "Voluntary Movement", in E. Kandel and J. Schwartz, (eds.), *Principles of Neural Science* . New York, NY: Elsevier, pp. 487-500.
Ghez, C. and Fahn, S. (1985) "The Cerebellum", in E. Kandel and J. Schwartz, (eds.), *Principles of Neural Science.* New York, NY: Elsevier, pp. 487-500.
Gibson, J. (1979), *The Ecological Approach to Visual Perception.* Boston, MA: Houghton Mifflin.
Goldman, A. (1976) "Discrimination and Perceptual Knowledge", *Journal of Philosophy*, 73, pp. 771-91.
Greenwood, J. (1991) "Self-Knowledge: Looking in the Wrong Direction", *Behavior and Philosophy*, 19, 35-47.
Hatfield (1988) "Representation and Content in Some (Actual) Theories of Perception", *Studies in History and Philosophy of Science*, 19, pp. 175-214.
Haugeland, J. (1981), "Introduction to *Mind Design*", in J. Haugeland (ed.) *MindDesign.* Cambridge, MA: MIT Press.
Haugeland, J. (1993), "Mind Embedded and Embodied", *Proceedings of Mind and Cognition: An International Symposium, May 28-30, 1993.* Taipei, Taiwan: Academia Sinica.
Heil, J. (1987) "Are We Brains in a Vat? Top Philosopher Says 'No'", *Canadian Journal of Philosophy*, 17, pp. 427-36.
Heil, J. (1988), "Privileged Access", *Mind*, 97, pp. 238-51.
Heil, J. (1992) *The Nature of True Minds.* Cambridge: Cambridge University Press.
Horgan, T. (1991) "Actions, Reasons, and the Explanatory Role of Content", in McLaughlin (Ed.), pp. 73-101.
Horgan, T. (1993) "Superdupervenience", *Mind*, Y, pp. xxx-xxx.
Hornsby, J. (1980) *Action.* New York: Routledge and Kegan Paul
Hornsby, J. (1986), "Physicalist Thinking and Conceptions of Behavior", in P. Pettit and J. McDowell (eds.), *Subject, Thought, and Context.* Oxford: Clarendon Press, pp. 95-115.

Jacob, P. (1990) "Is There Such a Thing as Narrow Content", *Philosophical Studies.*

Kaelbling, L. (1993) *Learning In Embedded Systems.* Cambridge, MA: MIT Press.

Kaplan, D. (1978) "Demonstratives", in J. Almog, J. Perry, and H. Wettstein (eds.) *Themes From Kaplan.* Oxford: Oxford University Press.

Karmiloff-Smith, A. (1992) *Beyond Modularity: A Developmental Perspective on Cognitive Science.* Cambrige, MA: MIT Press.

Katz, J. (1991) *The Metaphysics of Meaning.* Cambridge, MA: MIT Press.

Kim, J. (1982), "Psychophysical Supervenience", *Philosophical Studies,* 41, 51-70.

Kim, J. (1984), "Concepts of Supervenience", *Philosophy and Phenomenological Research,* 45, pp. 153-76.

Kim, J. (1987) "'Strong' and 'Global' Supervenience Revisited", *Philosophy and Phenomenological Research,* 48, 315-26.

Kirsh, D. (1991) "Today the Earwig, Tomorrow Man?", *Artificial I ntelligence,* 47, pp. 161-84.

Kripke, S. (1972) "Naming and Necessity", in D. Davidson and G. Harmon (eds.) *Semantics of Natural Language.* Dordrecht, Netherlands: D. Reidel.

Kripke, S. (1980) *Naming and Necessity.* Oxford: Basil Blackwell.

Kobes, B. (1989) "Semantics and Psychological Prototypes", *Pacific Philosophical Quarterly.*

Lepore, E. and Loewer, B. (1986) "Solipsistic Semantics", in P. French, T. Uehling, and H. Wettstein (eds) *Midwest Studies in Philosophy, X.* Minneapolis, MN: University of Minnesota Press.

Loar, B. (1988), "Psychological Content and Social Content", in Grimm and Merrill (eds.) *Contents of Thought.* Tuscon, AZ: Arizona University Press, pp. 99-110.

Losonsky, M. (1995) "Embedded Systems vs. Individualism", *Minds and Machines,* 5, pp. 357-371.

Ludlow, P. (1995) "Externalism, Self-Knowledge, and the Prevalence of Slow-Switching", *Analysis,* 55.1, pp.45-9.

Ludwig, K. (1992) "Brains in a Vat, Subjectivity, and the Causal Theory of Reference", *Journal of Philosophical Research,* 27, pp. 313-345.

Lycan, W. (1987) *Consciousness.* Cambridge: Cambridge University Press.

Maloney, C. (1990) "The Right Stuff", *Synthese.*

Marr, D. (1982), *Vision.* San Francisco, CA: Freeman.

MacDonald, C. (1989) *Mind-Body Identity Theories.* London: Routledge and Kegan Paul.
McClammrock, R. (1995) *Existential Cognition.* Chicago, IL: University of Chicago Press.
McDowell, J. (1986) "Singular Thought and the Extent of Inner Space", in Pettit and McDowell (1986).
McLaughlin, B. (Ed.) (1991) *Dretske and His Critics.* Oxford: Basil Blackwell.
McKinsey, M. (1991) "Anti-Individualism and Privileged Access", *Analysis,* 51, pp. 9-16.
McKinsey, M. (1993) "Curing Folk Psychology of Arthritis", *Philosophical Studies.*
McKinsey, M. (1994) "Accepting the Consequences of Anti-Individualism", *Analysis,* 54, pp. 124-28.
Merton, P. (1973), "How Do We Control the Contractions of Our Muscles?", *Scientific American,* May: 30-37.
Millikan, R. (1984) *Language, Thought, and Other Biological Categories.* Cambridge, MA: MIT Press.
Millikan, R. (1991) "Speaking Up For Darwin", in B. Loewer and G. Rey (eds.) *Meaning and Mind: Fodor and His Critics.* Oxford: Basil Blackwell.
Millikan, R. (1994) "On Unclear and Indistinct Ideas", *Philosophical Persepctives.*
Morten, P. (1993) "Individualism and Vision Theory", *Philosophy of Science.*
Nagel, T. (1978) "What Is It Like To Be a Bat".
Nelkin, N. (1993) "Patterns", *Mind and Language.*
Nelkin, N. (1996) *Consciousness and the Origins of Thought.* Cambridge: Cambridge University Press.
Owens, J. (1987), "In Defense of a Different Doppelganger", *The Philosophical Review,* 96, 521-554.
Owens, J. (1991) "Cognitive Access and Semantic Puzzles", in C. Anderson and J. Owens (eds.) *Propositional Attitudes.* Stanford, CA: Stanford University Press.
Owens, J. (1993), "Content, Causation, and Psychophysical Supervenience", *Philosophy of Science*, 60: 242-261.
Patterson, S. (1991), "Individualism and Semantic Development", *Philosophy of Science,* 58, 15-35.
Pettit, P. and McDowell, J. (1986) *Subject, Thought, and Context.* Oxford: Clarendon Press.
Port, R. and van Gelder, T. (1995), *Mind as Motion.* Cambridge, MA: MIT Press.
Putnam, H. ([1973] 1981), "Reduction and the Nature of Psychology", reprinted in modified form in J. Haugeland (ed.) *Mind Design.*

(Originally published in*Cognition* 2: 131-146.) Cambridge, MA: MIT Press.

Putnam, H. (1975) "The Meaning of 'Meaning'", in Putnam, *Mind, Language, and Reality, Philosophical Papers, Vol. 2.* Cambridge: Cambridge University Press, pp. 215-271.

Putnam, H. (1981) *Reason, Truth, and History.* New York: Cambridge University Press.

Rey, G. (1983) "Concepts and Stereotypes", *Cognition,* 15, pp. 237-262.

Robb, D. (1997) "The Properties of Mental Causation", *The Philosophical Quarterly,* 47, pp. 178-194.

Russell, B. (1912) *The Problems of Philosophy.* Oxford: Oxford University Press.

Rutkowska, J. (1993) *The Computatinal Infant.* Brighton: Harvester.

Saidel, E. (1994), "Discussion: Content and Causal Powers", *Philosophy of Science,* 61, pp. 658-665.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World.* Princeton, NJ: Princeton University Press.

Searle, J. (1983) *Intentionality.* Cambridge: Cambridge University Press.

Segal, G. (1988) "Review: Subject, Thought, and Context", *Mind and Language,* Vol. 3, pp. 288-300.

Segal (1989) "Seeing What Is Not There", *The Philosophical Review,* 98, pp. 189-214.

Segal (1991) "Defense of a Reasonable Individualism", *Mind,* 100, pp. 485-494.

Shapiro, L. (1992) "Darwin and Disjunction: Foraging Theory and Univocal Assignments of Content", *PSA Vol. 1,* pp. 469-80.

Shapiro, L. (1993) "Content, Kinds, and Individualism in Marr's Theory of Vision", *The Philosophical Review,* 102, pp. 489-513.

Simon, H. (1969, 1982), *The Sciences of the Artificial.* Cambridge, MA: MIT Press.

Smith, P. (1984) ""Could We Be Brains in a Vat?", *Canadian Journal of Philosophy,* 14, pp. 115-23.

Sober, E. (1984) *The Nature of Selection.* Cambridge, MA: MIT Press.

Stalnaker, R. (1991) "Narrow Content", in C. Anderson and J. Owens (eds.) *Propositional Attitudes.* Stanford, CA: Stanford University Press.

Stampe, D. (1977) "Towards a Causal Theory of Linguistic Representation", *Midwest Studies in Philosophy,* pp. 81-102.

Stampe, D. (1987) "The Authority of Desire", *The Philosophical Review,* 96.

Stampe, D. (1988) "Need", *Australasian Journal of Philosophy.*

Stich, S. (1983), *From Folk Psychology to Cognitive Science.* Cambridge, MA: MIT Press.

Travis, C. (1995) "Critical Notice: Order Out of Messes", *Mind*, 104, pp. 133-144.

Tuomela, R. (1989), "Methodological Solipsism and Explanation in Psychology", *Philosophy of Science*, 56, 23-47.

Tymoczko, T (1989) "In Defense of Putnam's Brains", *Philosophical Studies*, 57, pp. 281-97.

Ullman, S. (1979) *The Interpretation of Visual Motion.* Cambridge, MA: MIT Press.

Walker, V. (1990), "In Defense of a Different Taxonomy: A Reply to Owens", *The Philosophical Review* 99, 425-431.

Warfield, T. (1992) "Privileged Self-Knowledge and Externalism Are Compatible", *Analysis*, 52, pp. 232-237.

Warfield, T. (forthcoming) "Knowledge of Content and Knowledge of the World", *Philosophy and Phenomenological Research.*

Williams, D. (1966) *The Principles of Empirical Realism.* Springfield: Charles Thomas.

Wilson, R. (1992), "Individualism, Causal Powers, and Explanation", *Philosophical Studies* 68, 103-139.

Wilson, R. (1993), "Against *A Priori* Arguments for Individualism", *Pacific Philosophical Quarterly* 74, 60-79.

Wilson, R. (1994a), "Causal Depth, Theoretical Appropriateness, and Individualism in Psychology", *Philosophy of Science* 61, 55-75.

Wilson, R. (1994b), "Wide Computationalism", *Mind* 103, 351-372.

Wilson, R. (forthcoming) *Cartesian Psychology and Physical Brains.* Cambridge: Cambridge University Press.

Woodfield, A. (1982) "On Specifying the Contents of Thoughts", in Woodfield (ed.) *Thought and Object.* Oxford: Clarendon Press.

Yablo, S. (1992) "Mental Causation", *The Philosophical Review*, 101, pp. 245-280.

INDEX

Anti-Individualism
  defined, 2
Adams, F., 18-19
Adams Family, 165-172

Bach, K., 20, 22, 230
Boghossian, P., 27-60, 75
Brooks, R., 209
Brueckner, A., 37-44, 67-71, 232, 233
Burge, T., 4, 6, 8, 9
  on externalism, 12-19, 21
  on self-knowledge, 28-57
  on skepticism, 80-82
  on vision theory, 84-130, 237, 240
  on mental causation, 131-135, 140, 144-150, 160, 162-174
Butler, K., 219, 228, 243, 244

Clark, A. and Chalmers, D., 210-212
Crane, T., 20, 22

Davies, M., 71-80, 86, 120-21, 124-125, 126, 129, 235, 240
Descartes, R., 4, 5, 27, 227
Dretske, F., 197-199, 227, 229-30

Egan, F., 88, 92-101, 115, 144-147
Externalism
  defined, 2

Falvey, K. and Owens, J., 26, 29-57, 67-71, 75-78
Fodor, J., 4, 19, 131, 133-135, 162-174

Haugeland, J., 8, 205-210, 218-220, 221
Heil, J., 57-58, 60-62, 136, 142-143
Individualism
  defined, 2, 86-92
Internalism
  defined, 2

Kaelbling, L., 213-15
Karmiloff-Smith, A., 215-16
Kim, J., 136, 184, 239
Kripke, S., 20

Loar, B., 20-21
Losonsky, M., 214-218
Ludlow, P., 33-34

Marr, D., 84-130
McClammrock, R., 146, 160-162, 220-22
McKinsey, M., 26, 232

Nelkin, N., 3, 66

Putnam, H., 6, 12-14, 64-80

Q-B Strategy, 66-80

Robb, D., 137-140, 239
Russell, B., 26
Rutkowska, J., 216-218

Saidel, E., 150-158
Segal, G., 85, 86, 101-3, 104-5, 106, 119-24, 126, 215
Self-Knowledge, 25-63
Shapiro, L., 84-5, 103-4, 105-19
Simon, H., 206
Skepticism
  Content, 27-63
  External World, 64-82
Standard Strategy, 35-50
Supervenience, 135-139
Systemic Anti-Individualism, 199-223

Toumela, R., 187-189, 199-205, 223
Tropes, 137-144

Ullman, S., 108-9

Warfield, T.29-31, 74
Wilson, R., 8, 144-147, 182-196, 212-13